Politechnika Wrocławska

DZIEDZINA: Dziedzina nauk inżynieryjno-technicznych

DYSCYPLINA: Informatyka techniczna i telekomunikacja

ROZPRAWA DOKTORSKA

Wykorzystanie metod personalizacji w zadaniu rozpoznawania treści humorystycznych

Mgr inż. Julita Bielaniewicz

Promotor / Promotorzy:
prof. dr hab. inż. Przemysław Kazienko

Promotor pomocniczy:
dr inż. Jan Kocoń

Słowa kluczowe: przetwarzanie języka naturalnego, personalizacja, rozpoznawanie humoru, transfer uczenia, chatgpt

WROCŁAW 2024

Niniejszą pracę pragnę zadedykować Kamilowi Kanclerzowi za bycie źródłem nieskończonej inspiracji do pracy oraz za stałą motywację do pokonywania swoich przeszkód w drodze do celu.

Za wsparcie pragnę serdecznie podziękować moim rodzicom, Ewie i Tomaszowi Bielaniewicz, siostrze, Marcie Pawelec, oraz siostrzenicy, Julii Pawelec.

Streszczenie

Przedmiotem niniejszej rozprawy doktorskiej jest problematyka personalizacji w zadaniu rozpoznawania treści humorystycznych. Celem pracy było opracowanie nowatorskich modeli przetwarzania języka naturalnego (*ang. natural language processing*, NLP), które uwzględniają zróżnicowane, indywidualne poczucie humoru użytkowników. Dotychczasowe podejścia w rozpoznawaniu humoru koncentrowały się na uogólnionych modelach, co prowadziło do zaniedbywania subiektywnych odczuć odbiorców. W ramach pracy dokonano szczegółowego przeglądu teorii humoru oraz metod jego automatycznego rozpoznawania, wskazując na istniejące luki badawcze w zakresie personalizacji. Opracowane w rozprawie modele bazują na zaawansowanych architekturach sieci neuronowych, takich jak GRU oraz mechanizmy atencji, które pozwalają na precyzyjne uwzględnienie specyfiki użytkownika. Modele te zostały przetestowane na danych o różnym kontekście kulturowym i językowym, co umożliwiło weryfikację ich zdolności do dostosowywania się do indywidualnych preferencji w percepcji humoru. Dodatkowo, zastosowanie technik transferu wiedzy, a także zaawansowane badania nad procesem uczenia modeli na różnych zestawach danych przy wykorzystaniu wielokrotnych foldów, znacząco zwiększyło efektywność i elastyczność tych modeli. Takie podejście pozwoliło na dogłębne zbadanie zdolności modeli do generalizacji oraz adaptacji do nowych obszarów tematycznych, jednocześnie uwzględniając zróżnicowane, indywidualne preferencje użytkowników. Badania objęły również porównanie wyników generowanych przez model ChatGPT-3.5 z rezultatami uzyskiwanymi przez spersonalizowane modele predykcyjne. Chociaż model generatywny wykazał wysoką skuteczność, to spersonalizowane podejście okazało się bardziej adekwatne w kontekście rozpoznawania humoru zgodnie z indywidualnymi upodobaniami odbiorców. Rozprawa ta dostarcza istotnych dowodów na to, że personalizacja w zadaniu rozpoznawania treści humorystycznych znacząco poprawia jakość predykcji, wykraczając poza możliwości modeli uogólnionych. Wyniki pracy wskazują na konieczność indywidualizacji systemów NLP, co otwiera nowe horyzonty w zakresie analizy i generowania treści dopasowanych do specyficznych potrzeb i preferencji użytkowników, stanowiąc znaczący wkład w rozwój tej dziedziny.

Abstract

The subject of this doctoral dissertation is the issue of personalization in the task of recognizing humorous content. The aim of the study was to develop innovative natural language processing (NLP) models that take into account the diverse, individual sense of humor of users. Previous approaches to humor recognition focused on generalized models, which led to the neglect of subjective audience perceptions. As part of the research, a detailed review of humor theories and methods for its automatic recognition was conducted, identifying existing research gaps related to personalization. The models developed in this dissertation are based on advanced neural network architectures, such as GRU and attention mechanisms, which allow for precise consideration of user-specific traits. These models were tested on datasets with different cultural and linguistic contexts, enabling the verification of their ability to adapt to individual preferences in humor perception. Additionally, the application of knowledge transfer techniques, as well as advanced research on the training process of models using multiple data sets and fold-based methods, significantly improved the models' efficiency and flexibility. This approach allowed for an in-depth examination of the models' generalization abilities and their adaptation to new thematic domains, while still accounting for the diverse, individual preferences of users. The research also included a comparison of results generated by the ChatGPT-3.5 model with those obtained by personalized predictive models. Although the generative model demonstrated high effectiveness, the personalized approach proved to be more suitable for recognizing humor in line with the individual preferences of the audience. This dissertation provides substantial evidence that personalization in the task of humor recognition significantly improves prediction quality, surpassing the capabilities of generalized models. The findings emphasize the necessity for the individualization of NLP systems, which opens new horizons in the analysis and generation of content tailored to specific user needs and preferences, making a significant contribution to the development of this field.

SPIS TREŚCI

SPIS RYSUNKÓW

SPIS TABLIC

WYKAZ SKRÓTÓW I OZNACZEŃ

NLP Natural Language Processing

ML Machine Learning

DNN Deep Neural Networks

RNN Recurrent Neural Network

LSTM Long Short-Term Memory

GRU Gated Recurrent Unit

FC Fully Connected Layer

MSE Mean Squared Error

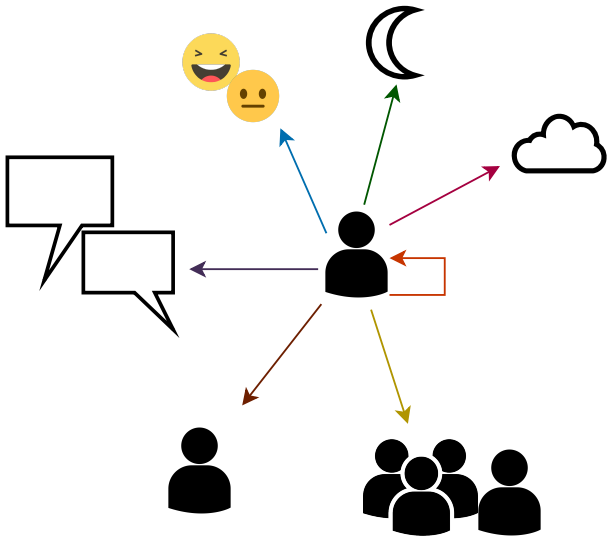
Część I

WPROWADZENIE

1 WSTĘP

Śmiech to jedna z najbardziej pierwotnych reakcji człowieka, jaką możemy wyrazić podczas wzajemnej komunikacji. Niezależnie od tego, czy jego źródło tkwi w celowym rozśmieszeniu kogoś, czy w wyniku niezamierzonego przypadku, jedno pozostaje niezmienne – nasze poczucie humoru jest subiektywne. Podobieństwo rozpoznawania śmieszności wśród ludzi istnieje pod względem bardziej powszechnej reakcji na wybrane fragmenty tekstu w określonych grupach społecznych lub kulturach. Jednak niektóre osoby mogą odbierać pewne treści za zabawne, podczas gdy inne ogólnie postrzegają je jako dziwne lub nudne. Często nietypowe, odbiegające od normy postrzeganie tekstu jest ignorowane. Zarówno uogólnienie, jak i subiektywność skupiają się na samym tekście, w tym na znaczeniu tych danych. Kiedy jednak próbujemy zmierzyć poziom śmieszności, obraźliwości lub zestawienia danych humorystycznych, staje się to zadaniem całkowicie zależnym od konkretnego człowieka i jego własnej perspektywy interpretacji śmieszności (Rys. 3). Obie te perspektywy to tylo część możliwych kontekstów interpretacji humoru, poza nimi istnieje też skoncentrowanie na ocenie ogólnej śmieszności zniezwiązanej z własnymi odczuciami, oceną samego siebie czy oceną tego co uważam jako część społeczności, a także co uważam na temat odczuć społeczności nie będąc jej częścią. Co więcej, nie tylko perspektywa odbioru treści jest czynnikiem wpływającym na odbiór treści humorystycznych. Można dodać szerszy kontekst, czasu, nastroju, dialogu, wątku, sytuacji (Rys. 1).

Bez względu na dziedzinę badawczą, jaką można się zajmować, w obliczu dużej ilości informacji i danych stosujemy różne sposoby na wydobycie pewnej wiedzy. Aby zbadać zależności, wykryć trendy oraz wyciągnąć wnioski z uzyskanych danych czy wyników eksperymentów, korzystamy z powszechnie stosowanego podejścia: generalizacji. To podejście pozwala dostrzec istotne wzorce w posiadanych danych, ale niesie ze sobą pewien koszt. Skupiając się na większości, pomijamy mniejszość. Jej jedynym zastosowaniem staje się podkreślenie większości, a dalszy zysk informacyjny nie istnieje lub okazuje się szumem, który zaburza hipotezy badawcze. Często pomijamy informację, jaką niosa za sobą wyniki odstające, mniejszościowe. Ten nieodkryty potencjał staje się szczególnie zauważalny w przypadku zadań, które opierają się na pewnej perspektywie, opinii lub odczuciu danej osoby. Do tej kategorii należy np. poczucie humoru, czyli określanie pewnego rodzaju kształtu śmieszności danego człowieka.



Rysunek 1: Przedstawienie tego, co może mieć wpływ na człowieka poprzez różne perspektywy postrzegania śmieszności. Jako odbiorca możemy oceniać komunikat subiektywnie (czerwony) ale też to, jak w naszej ocenie jest postrzegane przez nadawcę (fioletowy), przez danego odbiorcę (brązowy) czy daną grupę odbiorców (żółty). Ponadto, na ocenę mogą wpłynąć równiez czynniki takie jak nastrój (niebieski), pora dnia (zielony) czy pogoda (różowy). (Źródło: Opracowanie własne).

W kontekście badań nad przetwarzaniem języka naturalnego, szczególnie w odniesieniu do oceny humoru, jest obecny widoczny brak aspektu personalizacji. Poczucie humoru jest wysoce subiektywne, a różnice w jego odbiorze są na tyle znaczące, że trudno jest znaleźć dwie osoby o identycznym poczuciu humoru. Dlatego, aby badania były kompletne i uzupełnione o tę dodatkową perspektywę, konieczne jest uwzględnienie różnorodności indywidualnych reakcji na humor (Rys. 1).

Definiując subiektywność, można podać przykłady, takie jak ulubiony film, poczucie humoru czy najlepszy język programowania. Różnorodność odpowiedzi, które mogą się pojawić, jest nieskończona, ponieważ nie ma ludzi o identycznych opiniach ani jednego poprawnego wyboru w żadnej z wymienionych dziedzin. Podejście stosowane regularnie, nie tylko w przetwarzaniu języka naturalnego, ale ogólnie w badaniach, zapewnia nam rozwiązanie, które wybiera faworyta poprzez głosowanie większościowe. Metoda ta wydaje się być najbardziej intuicyjna, ale co z grupami mniejszościowymi, które mogły wybrać inne opcje? Metoda rankingowa odcina mniej popularne opcje, jednocześnie eliminując anotacje mniejszości.

Każdy człowiek, niezależnie od swojego pochodzenia kulturowego, ma swój subiektywny pogląd na kwestie takie jak odbiór emocjonalny, wrażliwość na treść czy poczucie humoru. W dziedzinie przetwarzania języka naturalnego perspektywa użytkownika jest jednym z najbardziej fascynujących i wymagających kierunków. Jest

ważnym elementem komunikacji i wytworem ludzkiej mądrości i kreatywności. Wraz z rozwojem sztucznej inteligencji konieczne jest, aby komputery analizowały i rozumiały humor tak jak ludzie, w celu poprawy inteligencji językowej komputerów i wydajności interakcji człowiek-komputer. Kiedy komputery wchodzą w interakcje z ludźmi, jeśli mogą modelować humor w ludzkim języku, a tym samym dokładnie identyfikować i rozumieć humorystyczne znaczenie języka, sprawi to, że komunikacja będzie bardziej humanitarna. Badania w dziedzinie sztucznej inteligencji poświęcone są wydobywaniu wzorców ludzkich zachowań i kategorii poznawczych, dlatego tak ważne dla rozwoju sztucznej inteligencji jest wykorzystanie komputerów do modelowania humoru i wydobywania jego istoty.

Ostatnie lata pokazały, że subiektywnemu obszarowi przetwarzania języka naturalnego poświęca się stopniowo coraz więcej uwagi, choć wciąż stosunkowo mniej niż dominujące podejście uogólniające. Jest to szczególnie widoczne, gdy koncentrujemy się na pojedynczym zadaniu personalizacji, takim jak wykrywanie emocji, mowa nienawiści lub zabawa. Choć pośrednio sygnalizuje to różnicę w ilości badań, możliwości, jakie tkwią w spersonalizowanym wykorzystaniu danych nawet w ramach pojedynczego subiektywnego zadania, są bardzo obiecujące (Kanclerz i in., 2020, 2021). Pośrednią kwestią, którą zwykle się pomija, jest natura grup i podgrup, które można wyróżnić na podstawie ogólnych cech poczucia humoru jednostek. Grupy te można posortować według wieku, płci, poglądów politycznych, zawodu i wielu innych cech (Hofmann i in., 2020). Niewątpliwie pozwoli to na dokładniejszy obraz poczucia humoru w poszczególnych grupach, ale nadal pozostanie ogólnie uogólnione. Istotne jest to, że w każdej grupie nadal będziemy obserwować wyjątkowe jednostki, które ostatecznie zostaną pominięte w większości. Niezależnie od tego, podążając tą ścieżką, moglibyśmy stworzyć w ramach wspomnianych grup podgrupy, które będą dokładniej sortować użytkowników, zachowując jednocześnie podejście generalizujące. Co więcej, prowadziłoby to do powstawania coraz większej liczby podgrup, które na nowo ustalałyby złoty standard dla pojedynczej grupy, wciąż nosząc uogólnienia, dopóki nie skupimy się na każdej osobie z osobna. Ten końcowy produkt wyznacza standardy poczucia humoru na osobę, co w istocie oznacza personalizację humoru. W ten sposób eliminujemy możliwość zignorowania osoby o nietypowym poczuciu humoru. Biorąc pod uwagę perspektywę kryjącą się w personalizacji (Kazienko i in., 2023; Kocoń i in., 2023, 2021b; Mieleszczenko-Kowszewicz i in., 2023), zaistniała potrzeba scenariuszy, w których istniałaby możliwość sprawdzenia różnych kombinacji podmiotowości i jej wpływu na działanie modelu. Zamierzaliśmy wdrożyć fuzję danych, która mogłaby poprawić ogólną wydajność spersonalizowanego rozumowania, pokazując, która kombinacja stosunku subiektywności do uogólnienia jest optymalna. Nasze badania zaowocowały odkryciem głównych korzyści płynących z modeli trenowanych na w pełni subiektywnym, połączonym zbiorze danych,

a także zauważeniem, że skutecznie zwiększa to jakość spersonalizowanych architektur głębokiego uczenia się. To osiągnięcie jest szczególnie ekscytujące, ponieważ otwiera drzwi do możliwości, jakie kryje się w dziedzinie fuzji danych.

Modele generatywne stanowią kluczową technologię w dziedzinie sztucznej inteligencji, szczególnie w dziedzinie przetwarzania języka naturalnego. Ich zadaniem jest tworzenie nowych treści na podstawie wzorców wyuczonych z dużych zbiorów danych. W odróżnieniu od tradycyjnych modeli predykcyjnych, modele generatywne nie tylko analizują istniejące dane, ale również potrafią generować nowe zdania, które odzwierciedlają strukturę i sens oryginalnych danych. Bazują najczęściej na architekturze transformera, co pozwala im na efektywne przetwarzanie długich sekwencji tekstu, uwzględniając złożone relacje między słowami.

Jednym z najnowocześniejszych przykładów takich modeli jest ChatGPT-3.5, który jest rozwinięciem technologii Generative Pretrained Transformer (GPT). ChatGPT-3.5 został przetrenowany na olbrzymich zbiorach danych tekstowych, co pozwala mu na prowadzenie rozmów w sposób naturalny i kontekstowy. Dzięki zdolności do przetwarzania wielkoskalowych danych i uczenia się z różnorodnych źródeł, ChatGPT-3.5 może generować teksty o wysokiej spójności i koherencji, dostosowując się do szerokiego zakresu tematów. Jednocześnie, jego architektura umożliwia bardziej zaawansowaną interakcję z użytkownikami, co czyni go jednym z czołowych narzędzi w dziedzinie NLP. Mając to na względzie, w ramach niniejszej pracy wykonano także badania zdolności modelu ChatGPT-3.5 do klasyfikacji (rozpoznawania) humoru dla dwóch wybranych zbiorów referencyjnych.

1.1 HUMOR JAKO KONSTRUKT PSYCHOLOGICZNY I JĘZYKOWY

Uczucie rozbawienia jest odczuwane przez każdą osobę inaczej. Zjawisko to ulega dalszemu wzmocnieniu w przypadku tekstów, gdzie ważną rolę odgrywa także dobór słów i ich różnorodne oddziaływanie na konkretnego użytkownika. Zarysowano zatem naturalną potrzebę modelowania indywidualnego poczucia humoru użytkownika.

Humor jest złożonym zjawiskiem, które można analizować zarówno z perspektywy psychologicznej, jak i językowej. Psychologiczne podejście do humoru koncentruje się na emocjonalnych i kognitywnych reakcjach wywołanych przez śmieszne bodźce. Humor może pełnić różne funkcje, takie jak poprawa nastroju, redukcja stresu, wzmacnianie więzi społecznych czy też demonstracja inteligencji i kreatywności. Klasyczne teorie psychologiczne, takie jak teoria sprzeczności, teoria nadwyżki energii czy teoria ulgi, starają się wyjaśnić mechanizmy, które prowadzą do tego, że coś jest odbierane jako zabawne.

Z językowego punktu widzenia, humor jest wynikiem specyficznego użycia języka, które wywołuje śmiech lub rozbawienie. Może to obejmować gry słów, ironię, sarkazm, absurdalne sytuacje czy nieoczekiwane zwroty akcji. Badania nad humorem językowym skupiają się na analizie struktur gramatycznych, semantycznych i pragmatycznych, które są charakterystyczne dla żartów i dowcipów. W kontekście przetwarzania języka naturalnego (NLP), analiza humoru staje się szczególnie trudna ze względu na jego subiektywny charakter oraz kontekstualne uwarunkowania.

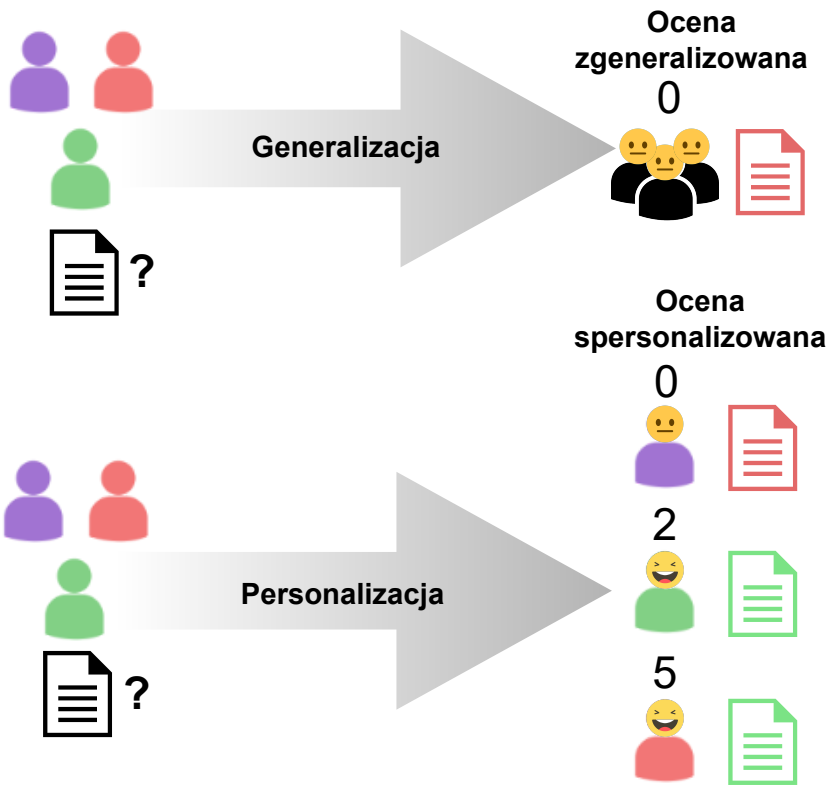
Zarówno psychologiczne, jak i językowe aspekty humoru wskazują na jego wielowymiarowość i złożoność. Zrozumienie, co sprawia, że coś jest zabawne, wymaga uwzględnienia indywidualnych różnic w odbiorze humoru, jak i specyfiki użytych środków językowych. Badania nad humorem w kontekście NLP muszą zatem uwzględniać personalizację, aby precyzyjnie ocenić śmieszność w sposób dostosowany do indywidualnych odbiorców.

1.2 MOTYWACJA I KONTEKST BADAŃ

Przegląd literatury (Rozdział 2) wykazał, że większość istniejących modeli NLP nie uwzględnia indywidualnych różnic w odbiorze humoru, co stanowiło punkt wyjścia do dalszych badań. Podczas poznawania wiedzy na temat poczucia humoru w kontekście przetwarzania języka naturalnego, występuje zauważalny brak prac w obszarze personalizacji, pomimo istotnej cechy poczucia humoru, jaką jest subiektywność. Nie istnieje bowiem jedno takie samo poczucie humoru, nie ma nawet grup ludzi, które podzielają swoje poczucie humoru w całości - każdy człowiek ma swoje własne, kształtowane od początku życia poczucie humoru. To, co jedną osobę rozśmiesza, może pozostawić inną zupełnie obojętną lub nawet zirytowaną. Ta subiektywność stwarza wyzwanie dla algorytmów NLP, które mają za zadanie ocenić lub generować treści humorystyczne. Standardowe modele NLP są zwykle trenowane na dużych zbiorach danych i dążą do uśrednionych wyników, co może prowadzić do utraty istotnych niuansów związanych z osobistymi preferencjami. Humor jest dziedziną zbadaną na pewnym poziomie, który nie ukazuje pełnego potencjału tej tematyki, a sama śmieszność nie jest analizowana w kontekście zrozumienia czym właściwie jest.

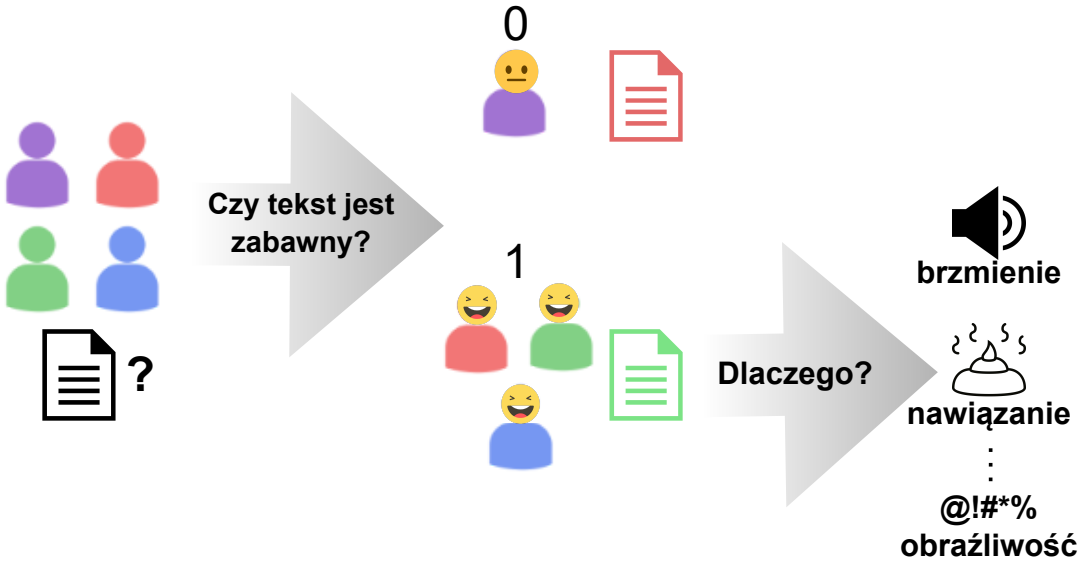
1.3 CEL I ZAKRES PRACY

Niniejsza praca doktorska ma na celu zgłębienie oraz dekompozycję zadania rozpoznawania śmieszności w dziedzinie przetwarzania języka naturalnego (NLP). Analiza literaturowa pozwoliła na zweryfikowanie obecnego stanu wiedzy w tej dziedzinie oraz zauważenie potrzeby rozszerzenia dotychczasowych badań o nowe rozwiązania,



Rysunek 2: Różnica między perspektywą uogólnioną (górną) i spersonalizowaną (dolną). W pierwszym dla każdego użytkownika dostępna jest ta sama prognoza. W drugim model generuje różne prognozy na podstawie indywidualnych preferencji użytkownika. (Źródło: (Bielaniewicz i in., 2022a)).

wnioski i efekty dogłębnych prac badawczych. rzeprowadzono szczegółowy przegląd literatury dotyczący istniejących metod rozpoznawania humoru w NLP, identyfikując zarówno osiągnięcia, jak i luki w badaniach. Analiza ta stanowiła podstawę do sformułowania celów i metod badawczych pracy. Opracowano nowe metody i algorytmy, które umożliwiają rozpoznawanie humoru z uwzględnieniem indywidualnych preferencji użytkowników. Personalizacja stanowi kluczowy element tych metod, odpowiadając na potrzebę bardziej precyzyjnego i adekwatnego oceniania śmieszności. W ramach pracy stworzono szczegółową kategoryzację rodzajów humoru, co pozwala na lepsze zrozumienie i modelowanie różnorodnych aspektów humorystycznych w kontekście NLP. Kategoryzacja ta obejmuje m.in. ironię, sarkazm, gry słowne, absurd i komedię sytuacyjną. Wykorzystano zaawansowane techniki, takie jak transfer uczenia, w celu wzmocnienia zdolności modeli uczenia maszynowego do rozpoznawania i generowania humoru. Transfer uczenia umożliwia przenoszenie informacji z jednego obszaru badawczego do drugiego, co zwiększa efektywność i precyzję modeli.



Rysunek 3: Proces oceny zabawności tekstu przy użyciu spersonalizowanych metod. Najpierw ustalamy, czy tekst jest *zabawny* (1) czy *nie* (0). Jeśli jest zabawny, uważamy go za jeden z sześciu rodzajów humoru. Ten sam tekst może być zabawny (z różnych powodów) lub nie, w zależności od osobistego poczucia humoru użytkownika. (Źródło: (Bielaniewicz i in., 2022a)).

Celem niniejszej pracy doktorskiej jest opracowanie i analiza metod dla zadania rozpoznawania śmieszności w przetwarzaniu języka naturalnego, z uwzględnieniem różnorodności i subiektywności humoru pojedyńczej osoby.

1.4 DEFINICJA PROBLEMU

Problem badawczy tej pracy doktorskiej polega na braku uwzględnienia personalizacji w istniejących modelach przetwarzania języka naturalnego (NLP) stosowanych do oceny humoru. Mimo że poczucie humoru jest z natury subiektywne i indywidualne, dotychczasowe badania i technologie NLP dążą do uśredniania wyników, co prowadzi do niedokładnych ocen i nieadekwatnych reakcji na treści humorystyczne. Aby skutecznie analizować i generować humor, konieczne jest opracowanie modeli NLP, które uwzględniają indywidualne różnice w poczuciu humoru, wynikające z unikalnych doświadczeń życiowych, kontekstów kulturowych i preferencji osobistych.

Celem niniejszej pracy jest zatem opracowanie i walidacja modeli NLP, które będą w stanie wnioskować w sposób spersonalizowany uwzględniając kontekst używtkownika. Jest to zrealizowane poprzez następujące zadania:

- Zebranie danych dotyczących indywidualnych preferencji humorystycznych.

- Opracowanie algorytmów (modeli), które uwzględniają te preferencje w procesie wnioskowania predykcji śmieszności.
- Testowanie i optymalizację modeli dla różnych konfiguracji.

Rozwiązanie tego problemu przyczyni się do lepszego zrozumienia, czym jest humor, oraz umożliwi tworzenie bardziej zaawansowanych i skutecznych systemów NLP, które będą mogły precyzyjnie oceniać i generować treści humorystyczne, uwzględniając unikalne preferencje i konteksty użytkowników.

1.5 PYTANIA / HIPOTEZY BADAWCZE

W ramach niniejszej pracy sformułowano następujące hipotezy badawcze:

H1. Zastosowanie metod personalizacji istotnie poprawi jakość predykcji metod rozpoznawania tekstów humorystycznych w porównaniu z podejściem wykorzystującym generalizację. (Podrozdział 7.6).

H2. Spersonalizowane architektury sieci neuronowych, w szczególności metody zaprezentowane w pracy, cechują się wyższą skutecznością w zadaniu rozpoznawania śmieszności w porównaniu do innym metod spersonalizowanych i do podejścia zgeneralizowanego (Podrozdział 7.6).

H3. Zaawansowane metody reprezentacji tekstu (architektury typu transformer) są lepsze niż proste metody reprezentacji (CBOW, Skipgram), które łącznie są lepsze niż metody niewykorzystujące wiedzy o tekście (Podrozdział 7.6.1).

H4. Metody spersonalizowanego rozpoznawania śmieszności cechują się lepszą jakością predykcji niż metody zgeneralizowane niezależnie od rozmiaru zbioru uczącego (Podrozdział 7.6.2)

H5. Dostarczenie wiedzy z innych zbiorów danych pomaga modelowi osiągać wyższą jakość wnioskowania w zadaniu wykrywania śmieszności (Podrozdział 7.6.3).

H6. Rodzaj treści tekstowych (słowotwórstwo, krótkie żarty, dłuższe wypowiedzi, nagłowki) wpływa na jakość predykcji modeli w zadaniu rozpoznawania śmieszności (Podrozdział 7.6.4).

H7. W uogólnionym zadaniu detekcji humoru generatywne modele ogólnego przeznaczenia (np. ChatGPT-3.5) mogą mieć mniejszą skuteczność niż dedykowane modele wytrenowane na danym zbiorze danych (Podrozdział 7.6.5).

1.6 ORYGINALNY WKŁAD PRACY

Niniejsza praca doktorska wnosi istotny wkład do dziedziny przetwarzania języka naturalnego (NLP) i analizy humoru poprzez kilka kluczowych osiągnięć:

1. Przeprowadzono kompleksowy przegląd literatury dotyczący istniejących badań nad humorem w kontekście NLP. Analiza ta obejmuje zarówno klasyczne teorie humoru, jak i nowoczesne podejścia technologiczne, identyfikując jednocześnie luki w badaniach, zwłaszcza w obszarze personalizacji (Rozdział 2).

2. Opracowano szczegółową kategoryzację rodzajów humoru, która uwzględnia różnorodność form humorystycznych, takich jak ironia, sarkazm, gry słowne, absurd czy komedia sytuacyjna. Ta kategoryzacja jest kluczowa dla zrozumienia i modelowania różnorodnych aspektów humoru w kontekście NLP (Podrozdział 3.4).

3. Stworzono innowacyjne architektury modelu uczenia maszynowego, specjalnie zaprojektowane by opierać się na spersonalizowanej predykcji śmieszności treści humorystycznych z uwzględnieniem indywidualnych preferencji (GRU oraz Wielogłowicowej Atencji). Modele te integrują różne techniki NLP i głębokiego uczenia, aby skutecznie personalizować predykcję śmieszności (Rozdział 6).

4. Przeanalizowano jakość wnioskowania dla spersonalizowanych architektur uczenia maszynowego i różnych modeli reprezentacji treści tekstowych oraz dla różnych zbiorów danych (w tym różnych języków i rodzajów humoru) dla zadania predykcji śmieszności. Uwzględniono zarówno nowe, zaproponowane w pracy architektury spersonalizowane jak i te, które opracowano w ramach grupy badawczej HumanNLP; wszystkie również względem referencyjnego rozwiązania zgeneralizowanego (Podrozdział 7.5.1 & 7.6.1, praca (Bielaniewicz i in., 2022a)).

5. Zbadano wpływ rozmiaru zbioru uczącego na jakość wnioskowania dla zadania predykcji śmieszności. Poprzez inkrementacyjne dodawanie foldów do zbioru uczącego opracowano szczegółową analizę tego, jak dobrze dany model uczenia maszynowego radził sobie z ograniczoną liczbą danych oraz tego, jak szybko jest w stanie zrozumieć charakterystykę i niuanse związane ze śmiesznością (Podrozdział 7.5.2 & 7.6.2).

6. Przeanalizowano jakość wnioskowania w scenariuszu transferu uczenia dla zadania predykcji śmieszności dla czterech różnych konfiguracji, obejmujących różny stopień ilości spersonalizowanych danych w zbiorze uczącym. Badano

jak wzbogacenie zbiorów danych o wiedzę spersonalizowaną jest w stanie pomóc w zrozumieniu charakterystyki i niuansów związanych z humorem (Podrozdział 7.5.3 & 7.6.3, praca (Bielaniewicz i in., 2022b)).

7. W ramach pracy w projekcie CLARIN przeprowadzono dwie procedury anotacyjne, których efektem są dwa zbiory danych w języku polskim: Doccano1 oraz Doccano2 zawierające anotacje 26 zadań, w tym zadań związanych z humorem, za które odpowiadałam. Zbiory te wraz z innymi dostępnymi publicznie były wykorzystane do testowania efektywności predykcji moich i referencyjnych architektur (Podrozdział 7.2.1, badania opublikowano w pracy (Bielaniewicz i in., 2022b)).

8. Przeprowadzono badania na modelu generatywnym ogólnego przeznaczenia ChatGPT-3.5 w celu porównania jego skuteczności względem modeli dedykowanymi uznawanych obecnie za najlepsze dla danego zadania (*SOTA: State-of-the-Art*). Zrealizowano to dla dwóch zadań zgeneralizowanej detekcji – klasyfikacji humoru (Podrozdział 7.5.4 & 7.6.5, praca (Kocoń i in., 2023)).

9. Według bazy Web of Science opublikowane prace, zawierające wyniki badań niniejszej pracy zostały zacytowane 153 razy, a według bazy Scopus prace zostały zacytowane 298 razy z wartością współczynnika h-index równą 6.

Wyniki prac zostały opublikowane w artykułach opublikowanych na międzynarodowych konferencjach naukowych i międzynarodowych recenzowanych czasopismach naukowych:

1. Kamil Kanclerz, Konrad Karanowski, **Julita Bielaniewicz**, Marcin Gruza, Piotr D. Miłkowski, Jan Kocoń, Przemysław Kazienko
   PALS: Personalized Active Learning for Subjective Tasks in NLP. W: The 2023 Conference on Empirical Methods in Natural Language Processing : Proceedings of the Conference : EMNLP 2023, December 6-10, 2023 / eds. Houda Bouamor, Juan Pino, Kalika Bali. Stroudsburg : Association for Computational Linguistics, cop. 2023. s. 13326-13341. (Kanclerz i in., 2023b)
   Ranga w Bazie CORE: A*, Lista ministerialna: 140 pkt.

2. **Julita Bielaniewicz**, Przemysław Kazienko
   From generalized laughter to personalized chuckles: unleashing the power of data fusion in subjective humor detection. W: 23nd IEEE International Conference on Data Mining Workshops (ICDMW), 1–4 December 2023 Shanghai, China : proceedings / eds. Jihe Wang [i in.]. Piscataway, NJ : Institute of Electrical and Electronics Engineers, cop. 2023. s. 716-725.. (Bielaniewicz i in., 2022b)
   Ranga w Bazie CORE: A*, Lista ministerialna: 200 pkt.

3. Kamil Kanclerz, **Julita Bielaniewicz**, Marcin Gruza, Jan Kocoń, Stanisław J. Woźniak, Przemysław Kazienko
   Towards model-based data acquisition for subjective multi-task NLP problems. W: 23nd IEEE International Conference on Data Mining Workshops (ICDMW), 1–4 December 2023 Shanghai, China : proceedings / eds. Jihe Wang [i in.]. Piscataway, NJ : Institute of Electrical and Electronics Engineers, cop. 2023. s. 726-735. (Kanclerz i in., 2023a)
   Ranga w Bazie CORE: A*, Lista ministerialna: 200 pkt.

4. Wiktoria Mieleszczenko-Kowszewicz, Kamil Kanclerz, **Julita Bielaniewicz**, Marcin Ł. Oleksy, Marcin Gruza, Stanisław J. Woźniak, Ewa Dziecioł, Przemysław Kazienko, Jan Kocoń
   Capturing human perspectives in NLP: questionnaires, annotations, and biases. W: Proceedings of the 2nd Workshop on Perspectivist Approaches to NLP co-located with 26th European Conference on Artificial Intelligence (ECAI 2023), Kraków, Poland, September 30th, 2023 / Gavin Abercrombie [i in.]. [B.m. : b.w.], 2023. s. 1-21. (Mieleszczenko-Kowszewicz i in., 2023)
   Ranga w Bazie CORE: A, Lista ministerialna: 140 pkt.

5. Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, **Julita Bielaniewicz**, Marcin Gruza, Arkadiusz Janz,Kamil Kanclerz, Anna A. Kocoń, Bartłomiej Koptyra, Wiktoria Mieleszczenko-Kowszewicz, Piotr D. Miłkowski, Marcin Ł. Oleksy, Maciej Piasecki, Łukasz Radliński, Konrad D. Wojtasik, Stanisław J. Woźniak, Przemysław Kazienko
   ChatGPT: Jack of all trades, master of none. Information Fusion. 2023, vol. 99, nr 101861, s. 1-37. (Kocoń i in., 2023)
   Ranga w Bazie CORE: A*, Lista ministerialna: 200 pkt., Impact Factor: 14,7

6. Przemysław Kazienko, **Julita Bielaniewicz**, Marcin Gruza, Kamil Kanclerz, Konrad Karanowski, Piotr D. Miłkowski, Jan Kocoń
   Human-centered neural reasoning for subjective content processing: hate speech, emotions, and humor. Information Fusion. 2023, vol. 94, s. 43-65. (Kazienko i in., 2023)
   Ranga w Bazie CORE: A*, Lista ministerialna: 200 pkt., Impact Factor: 14,7

7. **Julita Bielaniewicz**, Kamil Kanclerz, Piotr D. Miłkowski, Marcin Gruza, Konrad Karanowski, Przemysław Kazienko, Jan Kocoń
   Deep-SHEEP: sense of humor extraction from embeddings in the personalized context. W: 22nd IEEE International Conference on Data Mining Workshops (ICDMW), 28 November - 1 December 2022, Orlando, Florida : proceedings /

eds. K. Selçuk Candan [i in.]. Piscataway, NJ : Institute of Electrical and Electronics Engineers, cop. 2022. s. 967-974. (Bielaniewicz i in., 2022a)
Ranga w Bazie CORE: A*, Lista ministerialna: 200 pkt.

8. Kamil Kanclerz, Marcin Gruza, Konrad Karanowski, **Julita Bielaniewicz**, Piotr D. Miłkowski, Jan Kocoń, Przemysław Kazienko
What if ground truth is subjective? Personalized deep neural hate speech detection. W: LREC 2022 : Workshop Language Resources and Evaluation Conference : 20th June 2022, 1st Workshop on Perspectivist Approaches to NLP (NLPerspectives) : proceedings / eds. Gavin Abercrombie [i in.]. Paris : European Language Resources Association, cop. 2022. s. 37-45. (Kanclerz i in., 2022)
Ranga w Bazie CORE: C, Lista ministerialna: 20 pkt.

9. Jan Kocoń, Marcin Gruza, **Julita Bielaniewicz**, Damian Grimling, Kamil Kanclerz, Piotr D. Miłkowski, Przemysław Kazienko
Learning personal human biases and representations for subjective tasks in natural language processing. W: 21st IEEE International Conference on Data Mining ICDM 2021, 7-10 December 2021, Virtual Conference : proceedings / eds. James Bailey [i in.]. Piscataway, NJ : Institute of Electrical and Electronics Engineers, cop. 2021. s. 1168-1173. (Kocoń i in., 2021b)
Ranga w Bazie CORE: A*, Lista ministerialna: 200 pkt.

10. **Julita Bielaniewicz**, Adrianna Kozierkiewicz
A hybrid method of facial expressions recognition in the pictures. W: Computational Collective Intelligence : 12th International Conference, ICCCI 2020, Da Nang, Vietnam, November 30 - December 3, 2020 : proceedings / eds. Ngoc Thanh Nguyen [i in.]. Cham : Springer, cop. 2020. s. 581-590. (Lecture Notes in Computer Science. Lecture Notes in Artificial Intelligence, ISSN 0302-9743; vol. 12496) (Bielaniewicz i Kozierkiewicz, 2020)
Ranga w Bazie CORE: C, Lista ministerialna: 20 pkt.

1.7 UKŁAD PRACY

Niniejsza rozprawa doktorska składa się z czterech części, w których znajduje się dziewięć rozdziałów poświęconych różnym aspektom badań dotyczącym zadania śmieszności w dziedzinie przetwarzania języka naturalnego. Pierwsze trzy rozdziały znajdują sie w części pracy, zatytułowanej **Wprowadzenie**.

W **Rozdziale 1. Wstęp** przedstawiono teoretyczne i praktyczne podstawy badania humoru. Opisano humor jako konstrukt psychologiczny i językowy, a także nakreślono motywację oraz kontekst prowadzonych badań. Wyznaczono cel i zakres pracy,

zdefiniowano problem badawczy, sformułowano hipotezy oraz wskazano oryginalny wkład pracy, na końcu przedstawiając układ całej pracy.

Rozdział 2. Przegląd literatury omawia istniejące badania związane z poczuciem humoru. Zaczęto od ujęcia humoru w kontekście nauk społecznych, a następnie przeanalizowano podejścia do maszynowego rozpoznawania humoru z perspektywy informatyki. Na zakończenie zsyntetyzowano kierunki badań, tworząc ogólny przegląd literatury w tej dziedzinie.

Rozdział 3. Podstawowe definicje humoru to rozdział, w którym przedstawiono kluczowe pojęcia związane z humorem. Zdefiniowano humor, omówiono teorie humoru, istniejące taksonomie oraz zaproponowano nową taksonomię humoru. Ponadto, rozwinięto koncepcję personalizacji humoru, co stanowi ważny element dalszych badań.

Nastepne trzy rozdziały zostały umieszczone w części drugiej o tytule **Metody**.

W **Rozdziale 4. Metody generalizacji w badaniach poczucia humoru - metoda referencyjna oraz ChatGPT** przeanalizowano istniejące metody uogólniania wyników badań nad humorem. Następnie przedstawiono referencyjną architekturę badawczą, która stanowi uogólniony model do analizy danych dotyczących humoru. Ponadto, w tym rozdziale opisano metodę wykorzystania modelu generatywnego ChatGPT-3.5 w zadaniu rozpoznawania treści humorystycznych.

Rozdział 5. Metody uwzględniające spersonalizowane poczucie humoru koncentruje się na istniejących podejściach do personalizacji humoru. Szczegółowo omówiono algorytmy OneHot, Hubi-Formula, Hubi-Simple oraz Hubi-Medium, które są stosowane do modelowania indywidualnych preferencji humorystycznych.

W **Rozdziale 6. Nowe metody uwzględniające spersonalizowane poczucie humoru** wprowadzono nowe metody analizy śmieszności, w tym architekturę opartą na GRU oraz na wielogłowicowej atencji, które pozwalają na bardziej zaawansowaną personalizację w zadaniu predykcji treści humorystycznych.

Ostatnie trzy rozdziały są częścią sekcji trzeciej zatytułowanej **Badania**.

Rozdział 7. Badania eksperymentalne opisuje proces przeprowadzania badań nad śmiesznością. Omówiono plan badań, zbiory danych, scenariusze badań oraz metody badawcze. Następnie przedstawiono wyniki badań, ich analizę oraz dyskusję nad uzyskanymi wynikami, z której wyciągnięto wnioski końcowe.

W **Rozdziale 8. Podsumowanie** przedstawiono ogólny przegląd wniosków wynikających z pracy, podsumowując jej główne osiągnięcia i wyniki.

Ostatni **Rozdział 9. Kierunki dalszych prac**, opisuje potencjalne obszary dalszych badań nad humorem oraz wskazuje kierunki, w jakich można rozwijać temat personalizacji i predykcji treści humorystycznych.

Ostatnia, czwarta część pracy zatytułowana **Załączniki** jest poświęcona załącznikom z dokładnymi wykresami wyników badań, które szczegółowo pokazują wartości stanowiące podstawy do wniosków pracy.

2 PRZEGLĄD LITERATURY

Humor, jako zjawisko społeczne i psychologiczne, odgrywa istotną rolę w interakcjach międzyludzkich oraz w codziennym funkcjonowaniu jednostek. Jego złożoność wymaga wieloaspektowego podejścia do analizy, co obejmuje zarówno kontekstualne, jak i funkcjonalne aspekty humoru. Badania nad humorystycznymi interakcjami w różnych środowiskach, takich jak miejsce pracy i edukacja, oraz nad funkcjami humoru, dostarczają cennych informacji na temat jego wpływu na relacje społeczne i efektywność komunikacyjną. Dodatkowo, zrozumienie roli humoru w codziennym życiu oraz jego funkcji w regulacji emocjonalnej pozwala na szersze zrozumienie jego znaczenia w różnych kontekstach. W obszarze technologii, automatyczne rozpoznawanie humoru stanowi wyzwanie, które wiąże się z identyfikacją i klasyfikacją humorystycznych treści w tekście. Wyzwania te obejmują zarówno aspekty językowe, jak i kulturowe, co wymaga zaawansowanych metod analizy tekstu i uczenia maszynowego. Rozpoznawanie humoru w tekstach wymaga uwzględnienia różnorodnych kategorii humoru oraz kontekstów, w jakich się pojawiają.

2.1 POCZUCIE HUMORU Z PERSPEKTYWY NAUK SPOŁECZNYCH

Humor jest ważnym elementem interakcji społecznych, a jego rola w społeczeństwie i kulturze była przedmiotem wielu badań w różnych dziedzinach. W kontekście społecznym humor może służyć do budowania więzi międzyludzkich, rozładowywania napięć oraz regulacji emocji. Badania wskazują, że humor ma kluczowe znaczenie w nawiązywaniu i podtrzymywaniu relacji interpersonalnych oraz w budowaniu atmosfery zaufania i współpracy, szczególnie w kontekstach zawodowych i edukacyjnych (Martin i Ford, 2018). Humor odgrywa także ważną rolę w radzeniu sobie ze stresem i negatywnymi emocjami, co czyni go adaptacyjnym narzędziem w trudnych sytuacjach życiowych (Proyer, 2013). Badania nad humorem mają długą tradycję i obejmują szeroki zakres dyscyplin. Humor odgrywa ważną rolę w komunikacji międzyludzkiej, kulturze oraz interakcjach społecznych, dlatego był analizowany przez filozofów, psychologów, socjologów i lingwistów. W latach 70. XX wieku do badań nad humorem dołączyło podejście lingwistyczne. Victor Raskin wprowadził semantyczną teorię skryptu humoru, która opisuje humor jako wynik niespójności między dwoma skryptami poznawczymi, aktywowanymi w danym kontekście (Raskin, 1979). Raskin zdefiniował skrypt jako strukturę semantyczną, zawierającą informacje o sytuacjach i zdarzeniach, które są powiązane ze sobą w umyśle. Humor powstaje, gdy

skrypty te są niespodziewanie zmieniane, co prowadzi do rozpoznania niezgodności. Ta teoria stała się fundamentem dalszych badań nad humorem, takich jak teoria generowania humoru zaproponowana przez Attardo (Attardo i Raskin, 1991) oraz ontologiczna teoria humoru rozwinięta przez Raskina (Raskin, Hempelmann i Taylor, 2009), która wprowadziła formalne struktury umożliwiające automatyczną analizę i generowanie humoru. Jedną z najstarszych teorii humoru jest teoria wyższości, której korzenie sięgają XVII wieku i prac Thomasa Hobbesa. Hobbes uważał, że śmiech wynika z odczucia wyższości wobec innych, gdy dostrzegamy ich słabości. Współczesne rozwinięcie tej teorii zostało przedstawione przez Duncana (Duncan, 1985), który badał, w jaki sposób humor może wzmacniać relacje w grupach społecznych poprzez wyrażanie władzy i dominacji nad innymi. W żartach satyrycznych, szczególnie tych o wydźwięku politycznym, często pojawia się ten element, gdy humor polega na wyśmiewaniu osób publicznych lub idei, co buduje poczucie wyższości u nadawcy żartu.

Innym podejściem do rozumienia humoru jest teoria ulgi. Sigmund Freud w swoich pracach, szczególnie w książce (Freud, 1971) stwierdził, że humor pełni rolę mechanizmu obronnego, pozwalając na rozładowanie emocjonalnego napięcia i zmniejszenie lęków. W jego ujęciu żarty, szczególnie te, które dotyczą tematów tabu, pozwalają na przekroczenie społecznych granic w sposób, który jest akceptowany i wywołuje śmiech zamiast dyskomfortu. Freudowska teoria została rozszerzona przez Shurcliffa (Shurcliff, 1968), który badał korelację między humorem a zmniejszaniem napięcia emocjonalnego w trudnych sytuacjach.

Trzecią dominującą teorią humoru jest teoria niespójności. Humor według tej koncepcji wynika z poczucia zaskoczenia lub rozbieżności między oczekiwaniami a rzeczywistością. Teoria ta, rozwijana przez Veale'a (Veale, 2004) oraz Kulkę (Kulka, 2007), zakłada, że humor pojawia się, gdy zderzają się ze sobą dwa niepasujące elementy, a umysł próbuje je zintegrować, co prowadzi do śmiechu. Przykładem mogą być kalambury, które bazują na podwójnym znaczeniu słów lub nagłych zmianach kontekstu w wypowiedziach. Teoria niespójności jest szeroko stosowana w analizie komedii i dowcipów, gdzie gra słów, nieoczekiwane zakończenia lub nagłe zmiany perspektywy są kluczowymi elementami wywołującymi śmiech.

Jednym z istotnych tematów w literaturze dotyczącej śmieszności jest różnica w postrzeganiu humoru między różnymi kulturami. Badania pokazują, że w kulturach zachodnich humor jest częściej postrzegany jako pozytywny element codziennego życia, natomiast w kulturach wschodnich, takich jak Chiny czy Japonia, humor bywa traktowany z większym dystansem, szczególnie w sytuacjach formalnych (Jiang i Hou, 2019). Z tego względu wschodnie społeczeństwa rzadziej używają humoru jako strategii radzenia sobie z problemami, co może wpływać na różnice w ich dobrostanie psychicznym w porównaniu do społeczeństw zachodnich (Chen i Martin, 2007).

W kontekście edukacyjnym humor okazał się skutecznym narzędziem wspomagającym proces uczenia się. Badania pokazują, że nauczyciele stosujący humor w swoich lekcjach poprawiają zaangażowanie uczniów i sprzyjają lepszemu przyswajaniu trudnych zagadnień. W literaturze zauważa się jednak, że humor musi być stosowany z wyczuciem, aby nie zakłócać procesu nauczania (Proyer, 2013). Warto również podkreślić, że humor pełni ważną rolę w kontekstach zawodowych, ułatwiając współpracę oraz poprawiając atmosferę w miejscach pracy, co podkreślają badania nad zachowaniem w grupach roboczych (Gervais i Wilson, 2005).

2.2 MASZYNOWE ROZPOZNAWANIE HUMORU Z PERSPEKTYWY INFORMATYKI

Maszynowe rozpoznawanie humoru jest dynamicznie rozwijającą się dziedziną, szczególnie z uwagi na wzrost możliwości modeli językowych i metod głębokiego uczenia. Modele te, w tym popularne transformatorowe architektury, takie jak BERT i RoBERTa, wprowadziły znaczące usprawnienia w detekcji humoru poprzez analizę długich sekwencji tekstowych i kontekstowych powiązań pomiędzy słowami. Modele te mogą zrozumieć subtelności języka, takie jak ironia, sarkazm czy gry słowne, co wcześniej było dużym wyzwaniem dla systemów opartych na prostych regułach co wyszczególniono w pracach takich jak (Devlin i in., 2019) oraz (Hossain i in., 2020).

Wraz z rozwojem sztucznej inteligencji i uczenia maszynowego, badania nad humorem zaczęły obejmować także obszar przetwarzania języka naturalnego (NLP). Komputery zaczęły być wykorzystywane zarówno do generowania humoru, jak i jego rozpoznawania. Pierwsze badania w tym zakresie koncentrowały się na tworzeniu humorystycznych treści. Stock i Strapparava opracowali system, który automatycznie generował humorystyczne akronimy, stosując odpowiednie szablony i zasoby leksykalne (Stock i Strapparava, 2003, 2006). System ten znajdował zastosowanie w projektach rozrywkowych, a jego dalszy rozwój pozwolił na bardziej zaawansowane operacje związane z tworzeniem humoru.

Rozpoznawanie humoru za pomocą komputerów jest znacznie trudniejsze. Humor, który jest z natury subtelny, zależy od kontekstu, gry słów, a często od wiedzy kulturowej i emocjonalnej. Jak wskazują autorzy pracy (Taylor i Mazlack, 2004), próby rozpoznawania humoru zaczęły się od prostszych przypadków, takich jak dowcipy „Puk-Puk", gdzie humor opierał się na prostej grze słów. Wykorzystali oni n-gramy tekstowe oraz teorię skryptu humoru Raskina, aby stworzyć heurystyczne modele, które automatycznie wykrywały grę słów.

Ważnym wyzwaniem w kontekście predykcyjnego modelowania humoru jest rozpoznanie sarkazmu, ironii i innych bardziej złożonych form językowych, które nie są łatwo wykrywalne za pomocą standardowych metod. Modele, takie jak GRU i Long Short-Term Memory (LSTM), zyskały popularność w tej dziedzinie, ponieważ umoż-

liwiają one analizę sekwencji czasowych, uwzględniając wcześniejsze i późniejsze elementy tekstu, co jest kluczowe w rozpoznawaniu humoru osadzonego w kontekście narracyjnym (Hochreiter i Schmidhuber, 1997).

BERT, jako jeden z najbardziej zaawansowanych modeli przetwarzania języka naturalnego, został z powodzeniem zastosowany w zadaniach związanych z rozpoznawaniem humoru, zwłaszcza w edytowanych nagłówkach wiadomości, gdzie niewielkie zmiany w strukturze mogą przekształcić poważny tekst w humorystyczny (Mao i Liu, 2019). Modele te są trenowane na dużych zbiorach danych, takich jak nagłówki wiadomości i tweety, co pozwala im rozpoznać złożone wzorce w tekście.

Inne podejścia skupiały się na specyficznych cechach językowych humoru. Kidden i Brun (Kiddon i Brun, 2011) wyróżnili dwie kategorie cech: leksykalne, które wyrażają treści eufemistyczne, oraz strukturalne, opisujące uniwersalne wzorce zdań, często występujące w humorze erotycznym. Badania nad rozpoznawaniem humoru były kontynuowane przez Mihalceę i Strapparavę (Mihalcea i Strapparava, 2005), którzy zdefiniowali takie cechy jak aliteracja, slang dla dorosłych oraz antonimy, które posłużyły jako podstawa do trenowania klasyfikatorów humoru. Ich prace udowodniły, że metoda obliczeniowa może skutecznie rozpoznawać specyficzne formy humoru.

Mihalcea i Pulman (Mihalcea i Pulman, 2007) przeanalizowali także cechy tekstów humorystycznych na poziomie biegunowości emocjonalnej i orientacji na człowieka. Wyniki ich badań wykazały, że humorystyczne teksty często charakteryzują się ujemną biegunowością oraz odnoszą się do ludzi lub sytuacji z nimi związanych. Badania te wskazują, że nawet na poziomie sentymentu i emocji, humor wykazuje specyficzne wzorce, które mogą być automatycznie identyfikowane.

Zhang i in. (Zhang i in., 2014) podjęli próbę zintegrowania różnych teorii humoru z przetwarzaniem języka naturalnego, tworząc system zdolny do rozróżniania humoru na podstawie cech językowych i afektywnych. Zidentyfikowali oni kluczowe elementy humoru, takie jak niespójność oraz gra słów, które były wykorzystane do trenowania modeli językowych.

Purandare i in. (Purandare i Litman, 2006) przeanalizowali natomiast cechy prozodyczne i językowe związane z humorem, co pozwoliło im na zastosowanie metod automatycznego rozpoznawania humoru w mowie. Ich badania wykazały, że prozodia humoru, czyli rytm, intonacja i akcentowanie, różni się znacząco od zwykłej mowy, co może być użyte jako dodatkowy wskaźnik w systemach rozpoznawania.

Duża część literatury skupia się na zastosowaniu zaawansowanych modeli językowych w rozpoznawaniu humoru. Modele te, takie jak RoBERTa, oferują bardziej precyzyjne wyniki dzięki swojej zdolności do uchwycenia kontekstowych powiązań między słowami i zdań. Badania pokazują, że RoBERTa znacznie poprawia skuteczność w zadaniach związanych z humorem, takich jak generowanie humorystycznych treści czy rozpoznawanie edytowanych wiadomości, co widać w pracach (Liu i in.,

2019) oraz (Hossain i in., 2020). Modele te korzystają z pre-treningu na ogromnych zbiorach danych, co pozwala im uchwycić złożoność języka na poziomie leksykalnym i syntaktycznym.

Wśród bardziej zaawansowanych technik predykcyjnych, stosowanych w detekcji humoru, znajdują się również mechanizmy uwagi (attention mechanisms) oraz sieci rekurencyjne (GRU), które umożliwiają modelowanie długoterminowych zależności w sekwencjach tekstowych znanych z pracy (Hochreiter i Schmidhuber, 1997). Modele te są szczególnie skuteczne w zadaniach wymagających rozumienia kontekstu i przewidywania dalszego ciągu interakcji, co ma zastosowanie w systemach rozpoznających humor w czasie rzeczywistym, takich jak analiza komentarzy do treści wideo (Yang i in., 2019).

2.3 PODSUMOWANIE

Ewolucja badań nad humorem od tradycyjnych podejść psychologicznych i socjologicznych po nowoczesne zastosowania w uczeniu maszynowym i przetwarzaniu języka naturalnego pokazuje, jak skomplikowanym i wieloaspektowym zjawiskiem jest humor. Różne teorie, takie jak teoria wyższości, ulgi czy niespójności, pozwalają na głębsze zrozumienie mechanizmów, które kryją się za rozbawieniem. Jednak największym wyzwaniem pozostaje dokładne rozpoznawanie humoru w kontekście uczenia maszynowego. Choć osiągnięto postępy w automatycznym rozpoznawaniu humoru, szczególnie dzięki wykorzystaniu cech językowych czy emocjonalnych, pełne zrozumienie i generowanie humoru wciąż pozostaje trudnym zadaniem, zwłaszcza w obliczu różnorodności kontekstów kulturowych, językowych i indywidualnych. Zauważalny jest tu również brak ujęcia różnych perspektyw odbioru śmieszności, a także nastawienie na jednolitą formę interpretacji humorystycznych treści. Pomimo postępów, w literaturze dotyczącej detekcji humoru wciąż brakuje badań nad personalizacją. Humor jest zjawiskiem subiektywnym, a różnice kulturowe i indywidualne preferencje wpływają na sposób, w jaki użytkownicy odbierają humor. Obecne systemy są często zaprojektowane w sposób uogólniony, co oznacza, że nie dostosowują się do preferencji indywidualnych użytkowników. Badania nad personalizacją w systemach NLP, zwłaszcza w kontekście humoru, są nadal w początkowych fazach rozwoju.

3 PODSTAWOWE DEFINICJE HUMORU

W literaturze przedmiotu humor definiowany jest na wiele sposobów, co odzwierciedla złożoność i wieloaspektowość tego fenomenu. Aby lepiej zrozumieć, czym jest humor i jak można go skutecznie analizować za pomocą narzędzi NLP, konieczne jest zidentyfikowanie i omówienie podstawowych terminów oraz koncepcji, które będą stanowiły podstawę dla dalszych rozważań. W niniejszym rozdziale przedstawione zostaną kluczowe definicje humoru oraz teorie, które dostarczają ram teoretycznych niezbędnych do zrozumienia mechanizmów humorystycznych w tekstach.

3.1 POJĘCIE HUMORU

W literaturze przedmiotu można spotkać liczne próby definiowania humoru, które różnią się w zależności od przyjętej perspektywy teoretycznej i metodologicznej. Każda z tych definicji kładzie nacisk na inne aspekty humoru – od sprzeczności poznawczej, przez przewagę społeczną, po mechanizmy obronne psychiki.

Jednym z istotnych podejść do humoru jest teoria sprzeczności poznawczej, którą zaproponował Arthur Koestler. W swojej pracy (Koestler, 1964) opisał humor jako efekt zderzenia dwóch odrębnych, zazwyczaj niekompatybilnych schematów myślowych. Kluczowe w tej koncepcji jest niespodziewane połączenie tych sprzecznych elementów, co wywołuje napięcie, rozładowywane następnie przez śmiech. Zaskoczenie wynikające z nagłego połączenia nieoczekiwanych elementów jest tutaj kluczowe dla powstania humoru.

Inny wymiar humoru przedstawia teoria psychoanalityczna Sigmunda Freuda, według której humor pełni funkcję mechanizmu obronnego. W książce „Witz und seine Beziehung zum Unbewussten" (Freud, 1971) opisuje humor jako narzędzie umożliwiające radzenie sobie z trudnymi emocjami i nieuświadomionymi lękami. Freud podkreśla, że humor pozwala na rozładowanie napięcia psychicznego, które w innych okolicznościach mogłoby wywołać stres lub niepokój. Jak zauważa: „Istota humoru leży w zdolności do odnalezienia w każdej sytuacji warunków, które pozwolą przekształcić ją w żart" (Freud, 1905, s. 56). W tym kontekście humor działa jako narzędzie ochrony przed wewnętrznymi konfliktami emocjonalnymi, a śmiech jest formą rozładowania nagromadzonego napięcia.

Kolejne podejście do definiowania humoru przedstawia Salvatore Attardo, który w swojej pracy nad teorią dowcipu i humorem zaznacza, że humor można rozumieć jako „formę komunikacji, która wywołuje pozytywną reakcję emocjonalną poprzez ła-

manie konwencji społecznych i oczekiwań" (Attardo, 1994). W ujęciu Attardo humor jest narzędziem pozwalającym na przekraczanie norm i granic, co z jednej strony wywołuje u odbiorcy zaskoczenie, a z drugiej – radość wynikającą z przekroczenia tych ograniczeń. Definicja ta podkreśla rolę humoru jako aktu twórczego, który opiera się na zabawie z konwencjami i przyjętymi normami społecznymi.

W związku z wielona możliwymi sposobami na definiowanie humoru na podstawie powyższych definicji proponuję by zdefiniować humor jako złożone zjawisko poznawczo-emocjonalne, które powstaje w wyniku nieoczywistego połączenia sprzecznych lub nieoczekiwanych elementów, które prowadzą do rozładowania napięcia poprzez śmiech. Humor może pełnić różnorodne funkcje, takie jak wyrażenie wyższości społecznej, mechanizm obronny przed trudnymi emocjami czy narzędzie do radzenia sobie z poznawczymi sprzecznościami.

3.2 TEORIE HUMORU

Teorie humoru stanowią fundamentalny element w rozumieniu tego, jak i dlaczego ludzie śmieją się w różnych kontekstach. Każda z tych teorii oferuje inne spojrzenie na mechanizmy, które prowadzą do powstania humoru, oraz na to, jakie funkcje pełni on w życiu jednostki i społeczeństwa. W tym rozdziale zostaną omówione główne teorie humoru, które kształtowały badania nad tym zjawiskiem, ukazując różnorodność podejść do zrozumienia, czym jest humor, jakie są jego źródła oraz jakie skutki niesie za sobą jego stosowanie.

3.2.1 Teoria wyższości

Teoria wyższości, mająca swoje korzenie w pracach takich filozofów jak Platon i Arystoteles, a następnie rozwinięta przez badaczy takich jak (Duncan, 1985), sugeruje, że humor może wynikać z poczucia wyższości nad innymi. W myśl tej teorii, śmiech pojawia się, gdy ktoś czuje się lepszy od innych, czy to pod względem moralnym, intelektualnym, czy społecznym. Przykłady tej formy humoru często można znaleźć w żartach i satyrze, które wyśmiewają ludzkie wady, niedoskonałości czy niepowodzenia. Duncan podkreśla, że humor wyższościowy jest często wykorzystywany jako narzędzie społecznego porządku, pozwalając jednostkom i grupom na utrzymanie lub wzmocnienie swojej pozycji poprzez deprecjonowanie innych. Jest to mechanizm, który może prowadzić do wykluczania jednostek lub grup społecznych, wzmacniając stereotypy i uprzedzenia, co z kolei może mieć daleko idące konsekwencje społeczne.

3.2.2 Teoria ulgi

Teoria ulgi, wywodząca się z psychoanalizy i rozwinięta przez takich badaczy jak Freud oraz później (Shurcliff, 1968), zakłada, że humor pełni funkcję mechanizmu zmniejszającego napięcie emocjonalne. W tej perspektywie, śmiech jest reakcją na sytuacje stresowe, pozwalając na rozładowanie napięcia psychicznego i emocjonalnego. Shurcliff w swoich badaniach wskazuje, że humor może działać jako wentyl bezpieczeństwa, umożliwiając wyrażenie tłumionych emocji w sposób akceptowalny społecznie. W kontekście tej teorii, humor jest widziany jako zdrowy sposób radzenia sobie z lękiem, stresem i frustracją, co czyni go nie tylko zjawiskiem rozrywkowym, ale także ważnym elementem psychologicznego dobrostanu jednostki.

3.2.3 Teoria niespójności

Teoria niespójności, promowana przez badaczy takich jak (Veale, 2004) i (Kulka, 2007), kładzie nacisk na rolę zaskoczenia i niespójności w wywoływaniu śmiechu. W myśl tej teorii, humor pojawia się wtedy, gdy istnieje rozbieżność między oczekiwaniami a rzeczywistością, co prowadzi do komicznego efektu. Niespójność może przybierać różne formy – od gry słów po sytuacje, w których pozornie logiczne wnioski okazują się absurdalne. Veale i Kulka wskazują, że kluczowym elementem tej teorii jest zdolność odbiorcy do dostrzeżenia i zrozumienia tej niespójności, co często wymaga specyficznej wiedzy lub doświadczenia kulturowego. W praktyce teoria ta wyjaśnia, dlaczego pewne formy humoru są trudne do przetłumaczenia lub zrozumienia w innych kontekstach kulturowych – to, co jest zabawne w jednym kręgu kulturowym, może być całkowicie niezrozumiałe lub wręcz obraźliwe w innym.

3.2.4 Teoria Hay

Teoria humoru rozwinięta przez Jennifer Hay (Hay, 1995), jest znaczącym wkładem w badania nad mechanizmami tworzenia i odbioru humoru. W tej teorii, humor analizowany jest z perspektywy lingwistycznej i społecznej, kładąc szczególny nacisk na zrozumienie, jak różne aspekty komunikacji werbalnej i interakcji społecznych wpływają na percepcję humoru. W swojej pracy podkreśla, że humor jest ściśle związany z kontekstem społecznym oraz normami kulturowymi. Dodatkowo analizuje, w jaki sposób różnice płciowe wpływają na preferencje humorystyczne i jakie role społeczne są związane z tworzeniem humoru. Hay zauważa, że humor często odzwierciedla i wzmacnia stereotypy płciowe, co może prowadzić do różnych form humorystycznych w zależności od kontekstu płciowego. Praca sugeruje, że to, co uznawane jest

za śmieszne, nie jest jedynie wynikiem abstrakcyjnych zasad, lecz jest mocno zakorzenione w społecznych interakcjach i oczekiwaniach. Humor może być używany jako narzędzie do podkreślenia lub kwestionowania norm społecznych, co z kolei wpływa na jego odbiór w różnych grupach kulturowych. Hay wskazuje, że humor jest skutecznym narzędziem komunikacji, które może ułatwiać interakcje społeczne, łagodzić napięcia i budować więzi. W szczególności w badaniach nad interakcjami językowymi Hay koncentruje się na tym, jak humor może być używany do negocjowania znaczeń, tworzenia relacji i manipulowania emocjami. Humor w tym kontekście jest nie tylko mechanizmem rozrywkowym, ale także istotnym elementem strategii komunikacyjnych. Teoria Hay jest zatem przykładem podejścia, które łączy analizę językową i społeczną, aby zrozumieć, jak humor działa w różnych kontekstach komunikacyjnych i jak jest kształtowany przez normy społeczne i kulturowe. Jest dużym krokiem w kierunku podejścia personalizacji w analizie i interpretacji poczucia humoru u ludzi.

3.3 ISTNIEJĄCE TAKSONOMIE HUMORU

Taksonomia humoru odnosi się do klasyfikacji różnych rodzajów humoru, które pojawiają się w różnych kontekstach społecznych i kulturowych. W literaturze przedmiotu wyróżniono różne kategorie humoru, które odzwierciedlają jego wieloaspektowość oraz różnorodność funkcji i zastosowań. W tym rozdziale przedstawione zostaną kluczowe typologie humoru, oparte na pracach w dziedzinie psychologii i socjologii. Kada z nich wnosi istotne informacje na temat sposobów klasyfikacji humoru i jego roli w rónych rodowiskach.

3.3.1 Humor w Miejscu Pracy

(Vinton, 1989) skoncentrował się na humorze w kontekście zawodowym, tworząc kategorię humoru, który odgrywa kluczową rolę w miejscu pracy. Wyróżnił on humor integracyjny, który wspiera budowanie relacji i współpracy w zespołach, oraz humor dezintegrowujący, który może prowadzić do konfliktów lub izolacji. Humor w miejscu pracy jest często stosowany jako narzędzie do rozładowania napięć, poprawy atmosfery i wzmacniania poczucia przynależności w grupie. Vinton wskazał również, że humor może mieć wpływ na percepcję hierarchii i ról w organizacji, zmieniając dynamikę relacji między pracownikami i przełożonymi.

3.3.2 Humor w Klasie

(Neuliep, 1991) rozszerzył analizę humoru na środowisko edukacyjne, wskazując na jego znaczenie w kontekście nauczania i uczenia się. W tej perspektywie wyróżnia się kilka rodzajów humoru, takich jak humor dydaktyczny, który jest używany przez nauczycieli do ułatwienia nauki i zaangażowania uczniów, oraz humor relacyjny, który pomaga w budowaniu pozytywnych relacji między nauczycielami a uczniami. Humor w klasie może również pełnić funkcję wyrównawczą, pomagając w redukcji dystansu między nauczycielem a uczniami oraz w tworzeniu atmosfery sprzyjającej otwartości i kreatywności.

3.3.3 Funkcje Humorystyczne

(Graham, Papa i Brooks, 1992) zidentyfikowali różne funkcje humoru, co pozwoliło na dalsze rozróżnienie typów humoru w różnych kontekstach społecznych. W ich klasyfikacji wyróżnia się trzy główne funkcje: poznawczą, emocjonalną i społeczną. Humor poznawczy obejmuje humor związany z kreatywnym myśleniem i rozwiązywaniem problemów. Humor emocjonalny odnosi się do roli humoru w regulowaniu nastrojów i redukcji stresu, natomiast humor społeczny koncentruje się na funkcji humoru w budowaniu więzi i zarządzaniu relacjami interpersonalnymi. Te funkcje stanowią podstawę dla dalszych badań nad różnymi aspektami humoru i jego zastosowaniami w interakcjach międzyludzkich.

3.3.4 Humor w Komunikacji

(Hay, 1995) opracowała klasyfikację humoru w kontekście komunikacji międzyludzkiej, wskazując na trzy główne kategorie funkcji humoru: przyjemność, kontrolę i solidarność. Humor jako przyjemność jest stosowany do rozładowania napięć i zwiększenia komfortu interakcji. Funkcja kontroli dotyczy użycia humoru jako narzędzia do wywierania wpływu na innych oraz do wyrażania władzy czy krytyki w subtelny sposób. Solidarność odnosi się do wykorzystania humoru do wzmacniania więzi społecznych oraz budowania tożsamości grupowej. Hay zwrócił uwagę na strategiczne wykorzystanie humoru w komunikacji i jego rolę w kształtowaniu relacji społecznych.

3.3.5 Samo-regulacja Emocji

(Leist i Müller, 2013) zbadali rolę humoru w samo-regulacji emocji, koncentrując się na jego funkcji w zarządzaniu stresem i emocjami. W tej klasyfikacji wyróżniają się humor autoregulacyjny, który pomaga jednostkom radzić sobie z trudnymi emocjami, oraz humor rozładowujący, który służy do redukcji napięcia i zwiększenia poczucia kontroli nad własnymi emocjami. Humor, zwłaszcza w formie autoironicznej, może umożliwiać jednostkom zdystansowanie się od problemów i przyczyniać się do ich lepszego przetwarzania. Leist i Müller podkreślili, że skuteczność humoru w samo-regulacji może różnić się w zależności od indywidualnych cech osobowościowych i stylów radzenia sobie ze stresem.

3.3.6 Humor w Konfrontacji ze Śmiercią

(Lambert South, Elton i Lietzenmayer, 2022) zbadali humor w kontekście ekstremalnym, takim jak konfrontacja ze śmiercią. W ich badaniach wyróżnia się humor egzystencjalny, który jest używany w obliczu poważnych zagrożeń i wyzwań życiowych, takich jak śmierć. Humor ten może pełnić funkcję obronną, pomagając ludziom radzić sobie z lękiem i stresem związanym z egzystencjalnymi obawami. Autorzy wskazali, że humor w takich kontekstach może również wspierać więzi rodzinne i towarzyskie, oferując formę pocieszenia i wsparcia w trudnych czasach. Ich badania pokazują, jak humor może pełnić istotną rolę w radzeniu sobie z najcięższymi wyzwaniami życiowymi.

3.4 PROPOZYCJA NOWEJ TAKSONOMII

Zauważając istotny brak uporządkowania w istniejących taksonomiach charakteryzujących to, dlaczego coś może śmieszyć, opracowano nową kategoryzację. Jest nią ustrukturyzowany układ taksonomii, która określa następujący porządek:

- **Intencja**
  - **Cel**
    - **Powód**

Intencja określa Zebrane powody śmieszności z dotychczasowych charakterystyk pomogły uzupełnić sekcję **Powód**.

- *Intended* - intencjonalna, tj. ze świadomą chęcią rozśmieszenia
  - *Self* - cel: z samego siebie

    * *Praise* - by pochwalić
    * *Deprecation* - by oczernić
    * *To impress* - by zaimponować
    * *To relieve stress* - by rozładować stres
    * *Humanization* - by uczłowieczyć
- *Others* - cel: z innych
    * *Praise* - by pochwalić
    * *Persuasive* - by przekonać
    * *Deprecatrion* - by zdeprecjonować
    * *Affirmative* - by wesprzeć
    * *Punishing* - by ukarać
        · *Neutral* - w sposób neutralny
        · *Aggressive* - w sposób agresywny
        · *Passive-aggressive* - w sposób pasywno-agresywny
    * *Teasing* - by się podroczyć, podrażnić
        · *Friendly* - w sposób przyjazny
        · *Unfriendly* - w sposób nieprzyjazny
    * *Manipulation* - by zmanipulować
    * *To investigate* - by coś zbadać, dowiedzieć się informacji
    * *Bullying* - by się znęcać
- *No direct target* - bez konkretnego celu
    * *Speaking ill of the dead* - by mówić negatywnie o zmarłych
    * *Explanation* - w celu wyjaśnienia
    * *Suicide humor* - wisielczy humor
- *Miscellaneous* - cel: różnorodny; grupa,społeczność
    * *Wordplay* - gła słów
    * *Criticism* - by skrytykować
        · *Vicious* - w sposób złośliwy
        · *Constructive* - w sposób konstruktywny
    * *Sound* - zabawny wydźwięk
    * *Kinesic* - związany z ruchem

        * *Euphemistic* - eufemistyczny
        * *Xenophobic* - ksenofobiczny
        * *Fecal* - humor toaletowy
        * *Exaggerated* - wyolbrzymiony
            · *With ill intent* - ze złośliwą intencją
        * *Vulgar* - wulgarny
        * *Innuendo* - dwuznaczny
        * *Inside-joke* - żart środowiskowy, rozumiany przez okresloną grupę
        * *Ironic* - ironiczny

- *Unintended* - nieintencjonalna, brak woli rozśmieszenia
  - *Awkward* - przez sytuację niezręczną
  - *Situational* - sytuacyjny, kontekstowy
  - *Group chaining* - dopowiadanie żartów przez grupę w ramach odpowiedzi na poprzedni
  - *Absurdities* - absurdalny
- *Either* - brak konkretnej woli rozbawienia, jest chęć zaangażowania w kontakt, w bycie wysłuchanym
  - *Observational* - obserwacyjny
  - *Offensive* - ofensywny, obrażający
  - *Morality* - poruszający tematy moralności
  - *Trauma processing* - w celu przetworzenia traumy
  - *Coping with tragedy* - w celu poradzenia sobie z tragedią

3.5 IDEA PERSONALIZACJI HUMORU

Istnieje wiele powodów, dla których pewne wiadomości nas śmieszą i sprawiają, że w odpowiedzi reagujemy na różne sposoby. Aby uszanować wyjątkowość uczuć i indywidualne poczucie humoru, którego nie da się podrobić i wykorzystać ten unikatowy element, któy posiada każda osoba, należy zwrócić uwagę na kluczowy aspekt interpretacji śmieszności, jakim jest subiektywność odbioru humoru.

Personalizacja humoru odnosi się do sposobu, w jaki humor jest dostosowywany do indywidualnych cech, preferencji i kontekstu osoby lub grupy. W praktyce oznacza to, że humor może być tworzony i odbierany w sposób, który uwzględnia unikalne

doświadczenia, przekonania oraz zainteresowania odbiorcy. Personalizacja humoru zwiększa jego skuteczność, ponieważ żarty i dowcipy są bardziej trafne i dostosowane do specyficznych cech jednostki. Z psychologicznego punktu widzenia, personalizacja humoru związana jest z teorią poznawczą, która podkreśla, że nasze postrzeganie i interpretacja humoru są głęboko zakorzenione w naszych osobistych schematach poznawczych i emocjonalnych. Ludzie mają różne zestawy doświadczeń, przekonań i emocji, które wpływają na to, co uznają za zabawne. Na przykład, osoby o podobnych doświadczeniach życiowych czy wspólnych wartościach są bardziej skłonne do reagowania na humor, który odnosi się do tych wspólnych elementów. Personalizacja humoru w tym kontekście zwiększa jego efektywność, ponieważ lepiej odpowiada na indywidualne oczekiwania i wrażliwość odbiorcy. Chociaż możemy ocenić fragment tekstu z naszej perspektywy, typową kategoryzację reprezentuje binarna klasa śmieszności przypisana do tekstu: *funny* (klasa 1) lub *not Funny* (klasa 0). W tym uogólnionym podejściu badacze trenują model, aby wnioskować o tej samej klasie lub poziomie humoru dla każdego użytkownika, (Rys. 2, góra). Takie podejście jest szeroko rozpowszechnione w rozpoznawaniu humoru, klasyfikacji (Yang i in., 2015) i regresji, tj. przewidywaniu średniej śmieszności zagregowanej na podstawie wszystkich zebranych anotacji (Castro i in., 2018; Hossain i in., 2020, 2019). Koncepcja spersonalizowana przyjmuje całkiem inne założenia. Zakładamy, że etykieta klasy czy poziom śmieszności zależy od użytkowników i ich poczucia humoru, a nie tylko od zdefiniowanego wcześniej stopniowania wśród tekstów oraz uśrednionych lub zagregowanych anotacji. W rezultacie spersonalizowany model rozumowania zapewnia każdemu użytkownikowi inny wynik, (Rys. 2). Biorąc pod uwagę niniejsze informacje zauważa się istotną potrzebę opracowania modelu spersonalizowanego, którego działanie należałoby porównać z modelem działającym na zasadzie uogólniającej w celu zweryfikowania jakości zaproponowanego rozwiązania.

Część II

METODY

4 METODY GENERALIZACJI W BADANIACH POCZUCIA HUMORU - METODA REFERENCYJNA ORAZ CHATGPT

W kontekście badań nad poczuciem humoru, metody generalizacji odgrywają kluczową rolę w rozwoju modeli uczenia maszynowego, które mogą skutecznie przewidywać i analizować humor w różnych kontekstach. Generalizacja w tym przypadku odnosi się do zdolności modelu do stosowania wzorców wyuczonych na danych treningowych do nowych, nieznanych danych. W niniejszym rozdziale omówimy, jak techniki uczenia maszynowego przyczyniają się do procesu generalizacji w badaniach nad poczuciem humoru, oraz jakie metody i podejścia są wykorzystywane w tym kontekście.

4.1 ISTNIEJĄCE METODY UOGÓLNIAJACE

W badaniach nad poczuciem humoru, klasyfikatory są często używane do rozpoznawania i klasyfikowania różnych form humoru. Klasyfikatory, takie jak maszyny wektorów nośnych (SVM), drzewa decyzyjne czy klasyfikatory Bayesa, mogą być szkolone na danych etykietowanych, aby rozróżniać między różnymi typami humoru, np. sarkazmem, ironią czy dowcipami. Aby zapewnić skuteczną generalizację, modele te muszą być trenowane na zróżnicowanych danych, które reprezentują szeroki zakres kontekstów i stylów humoru. Sieci neuronowe, zwłaszcza głębokie sieci neuronowe (DNN) i rekurencyjne sieci neuronowe (RNN), oferują zaawansowane techniki modelowania, które mogą uchwycić złożone wzorce w danych humorystycznych. Modele takie jak Long Short-Term Memory (LSTM) oraz transformery (np. BERT, GPT) są zdolne do analizy kontekstualnej i semantycznej, co jest istotne w rozpoznawaniu i generowaniu humoru. Generalizacja w tym kontekście jest osiągana poprzez stosowanie dużych zbiorów danych oraz technik regularyzacji, takich jak dropout czy normalizacja, które pomagają zapobiegać przeuczeniu (overfitting) i poprawiają zdolność modeli do generalizowania na nowe dane. Badania poświecone tematyce poczucia humoru w przetwarzaniu języka naturalnego charakteryzują się zastosowaniem osadzeń słów, które pozwoliły systemom wykrywania humoru ewoluować w bardziej złożone struktury. Skutecznie wychwytują one preferencje użytkowników w szerokim zakresie zadań NLP (Kocoń i in., 2021a; Tang i in., 2015), co jest powodem, dla którego są one również wykorzystywane w wielu najnowszych pracach dotyczących złożonych rozwiązań w systemach rekomendacji. Obiecujący potencjał tego podejścia może być przekonującym argumentem, dlaczego jest ono uważane za kluczowy ele-

ment możliwego przyszłego rozwoju subiektywności w NLP. Użytkownik może być również reprezentowany jako unikalna sekwencja tokenów z zasobu modelu (Mireshghallah i in., 2021). W tym przypadku zadaniem jest nauczenie modelu unikalnych osadzeń, które powinny uchwycić indywidualne preferencje użytkownika.

Kolejny istotny zakres metod przyczyniających się do rozwoju personalizacji w rozpoznawaniu humoru można zauważyć w systemach rekomendacji żartów, takich jak (Miao i in., 2016). Te podejścia w personalizacji humoru zapewniały wysokiej jakości rozumowanie poprzez zastosowanie modeli głębokiego uczenia i danych dotyczących humoru oraz jego źródła. Źródło to odnosi się do kontekstu zrozumiałego dla użytkowników, którzy następnie mogli zidentyfikować, dlaczego coś jest śmieszne (Amir i in., 2016; Chakrabarty i in., 2019).

Eksploracja interpretacji humoru przez użytkownika doprowadziła do szybkiego postępu w badaniach nad wyrażeniami humorystycznymi (Engelthaler i in., 2018; Westbury i in., b.d.). Zgeneralizowane oceny pozwoliły autorom uzyskać długą listę często spotykanych zabawnych słów. Jednak podejście to pomijało wielu użytkowników o niestandardowym poczuciu humoru. To stanowi motywację do zaprojektowania dedykowanych, spersonalizowanych metod wykrywania humoru. Uważa się, że jest to przełom w tej dziedzinie, ponieważ nie tylko chroni użytkowników, którzy zazwyczaj byliby traktowani jako anomalia ale też poraz pierwszy poświęca im uwagę i traktuje na równi z popularnymi poglądami użytkowników.

4.2 ARCHITEKTURA MODELU REFERENCYJNRGO (UOGÓLNIONA)

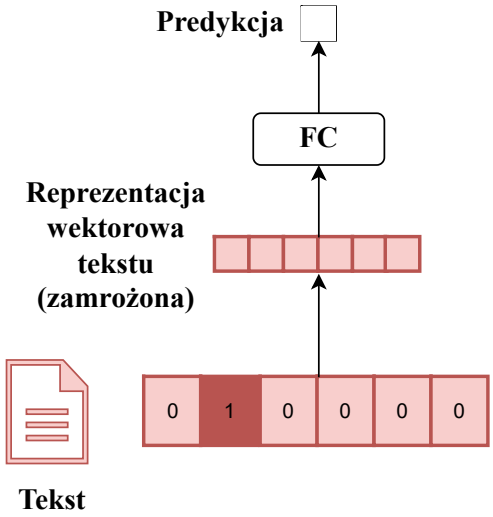
Architektura ta (Rys. 4), również nazywana architekturą *baseline*, reprezetuje podejście generalizacyjne, t.j. większościowe. Model Referencyjny został uwzględniony w celu zestawienia jakości predykcji modelu w porównaniu z modelami opierającymi się na podejściu spersonalizowanym. Architektura ta została opracowana w ramach wspólnej pracy w zespole badawczym HumaNLP i opublikowana w pracy (Kocoń i in., 2021b).

Na początku model rozpoczyna przetwarzanie od przyjęcia reprezentacji wektorowych tekstów jako danych wejściowych. Te osadzenia reprezentują tekst w postaci numerycznej i są wstępnie przetworzone.

Model zawiera pojedynczą warstwę liniową, która transformuje reprezentacje wektorowe tekstów o określonym wymiarze na przestrzeń o innej, ustalonej liczbie wymiarów, odpowiadającej liczbie wyjściowych kategorii. Warstwa ta działa jako klasyfikator, ucząc się mapować wektory wejściowe na odpowiednie kategorie wyjściowe.

Jest to prosta architektura działająca na zasadzie wnioskowania większościowego. Służy jako podstawowy standard, z którym porównuje się modele spersonalizowane. Dzięki temu można ocenić, w jakim stopniu dodatkowe warstwy, mechanizmy uwagi

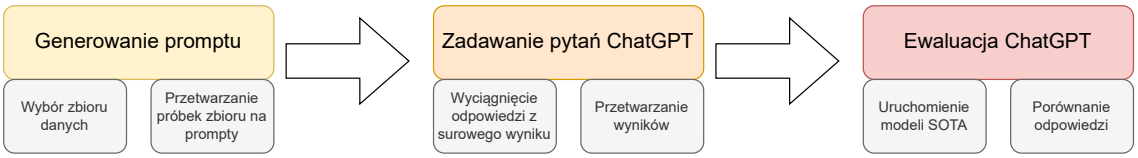
lub inne złożone komponenty związane z personalizacją przyczyniają się do poprawy wydajności w zadaniu klasyfikacji treści humorystycznych. Model referencyjny został opracowany w ramach pracy w zespole badawczym.



Rysunek 4: Model referencyjny - model bazowy opierający się wyłącznie na treści tekstowej, bez uwzględniania danych dotyczących użytkownika (bez personalizacji). (Źródło: (Kazienko i in., 2023)).

4.3 CHATGPT - UOGÓLNIONE WNIOSKOWANIE O HUMORZE

Przetestowano model ChatGPT w wersji 3.5 na zadaniach skupiających się na zgeneralizowanym humorze w dziedzinie przetwarzania języka naturalnego. Zadanie to obejmuje stosunkowo prostą binarną klasyfikację tekstów pod kątem śmieszności, ale także powiązane podzadania, takie jak sarkazm, ironia, gra słów czy inne. Uwzględniono również semantyczną anotację i akceptację tekstu w kontekście zrozumienia języka naturalnego (NLU), jak na przykład rozróżnianie znaczeń słów (WSD) i odpowiadanie na pytania na podstawie tekstu wejściowego. Przebieg ewaluacji prezentuje rysunek (Rys. 5).



Rysunek 5: Diagram ewaluacji ChatGPT pokazujący trzy etapy przetwarzania danych: 1) wybór zbioru danych i przekształcenie zbioru testowego w formę opartą na promptach; 2) zadawanie pytań (promptowanie) w usłudze ChatGPT; 3) wyodrębnianie etykiet z surowych wyników, ocena przy użyciu wartości referencyjnych oraz porównanie wyników z modelami SOTA. (Źródło: Opracowanie własne).

W celu zebrania informacji na temat oceny śmieszności zadanych tekstów ze zbioru ColBERT oraz Sarcasm, przetestowano teksty ze zbiorów i na podstawie treści w zadanym prompcie poproszono o ocenę zadanego tekstu. Poniżej znajdują się przykłady dla zbierania odpowiedzi od modelu ChatGPT3.5 dla obu tych zadań w strukturze (*prompt, odpowiedź, oczekiwana odpowiedź*):

- Zbiór: ColBERT
  - Prompt: *Który z atrybutów: śmieszne, nieśmieszne opisuje dany tekst?*
  - Tekst ze zbioru: *Każda konkurencja olimpijska powinna zawierać jedną przeciętną osobę rywalizującą jako punkt odniesienia.*
  - Odpowiedź ChatGPT: nieśmieszne
  - Oczekiwana odpowiedź: nieśmieszne
- Zbiór: Sarcasm
  - Prompt: *Który z atrybutów: śmieszne, nieśmieszne opisuje dany tekst?*
  - Tekst ze zbioru: *żaden problem, każdy może nosić trampki, ale nie każdy potrafi zapuścić tak solidny wąs*
  - Odpowiedź ChatGPT: śmieszne
  - Oczekiwana odpowiedź: śmieszne

5 METODY UWZGLĘDNIAJĄCE SPERSONALIZOWANE POCZUCIE HUMORU

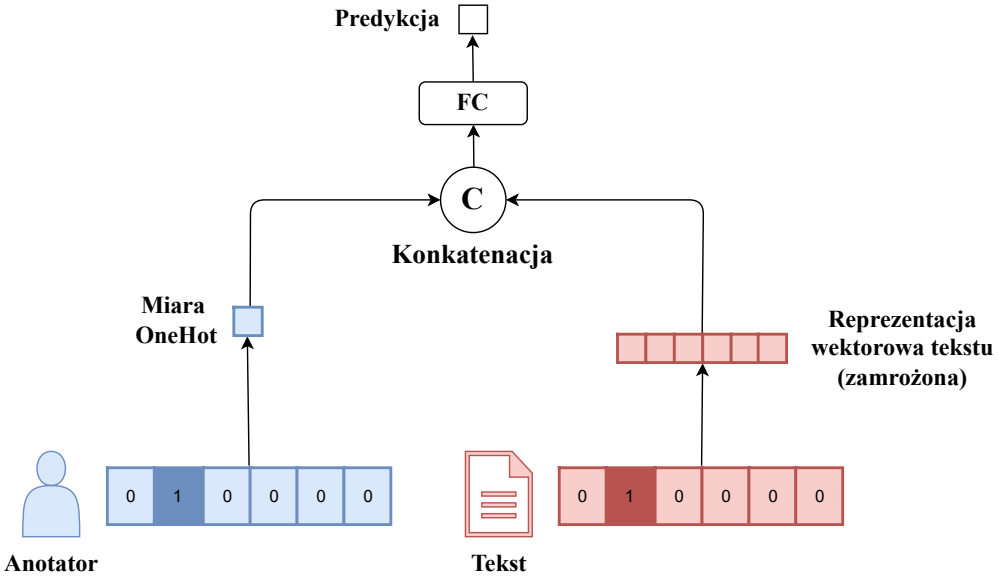
W kontekście uczenia maszynowego, uwzględnianie spersonalizowanego poczucia humoru stanowi zaawansowany obszar badań, który koncentruje się na dostosowywaniu modeli i algorytmów do indywidualnych preferencji i kontekstów humorystycznych użytkowników. Personalizacja humoru w systemach uczenia maszynowego wymaga nie tylko zrozumienia ogólnych wzorców humorystycznych, ale także umiejętności adaptacji do specyficznych cech użytkownika, takich jak jego zainteresowania, styl komunikacji oraz kontekst społeczny. Wprowadzenie metod personalizacji ma na celu poprawę trafności i efektywności rekomendacji humorystycznych, co jest kluczowe w aplikacjach takich jak chat-boty, platformy rozrywkowe oraz narzędzia analizy treści. W niniejszym rozdziale omówimy techniki i strategie stosowane w celu integracji spersonalizowanego poczucia humoru w modelach uczenia maszynowego, zwracając szczególną uwagę na wyzwania i innowacje w tej dziedzinie.

W podejściu spersonalizowanym zakłada się, że o poziomie śmieszności tekstu decydują użytkownicy i ich indywidualne poczucie humoru, a nie sam tekst i przeciętne lub łączone anotacje. Założenie to opiera się na indywidualnych preferencjach użytkownika, które należy wykorzystać, aby otrzymać odrębny wynik dla każdego użytkownika, w oparciu o jego indywidualne preferencje, jak pokazano na (Rys. 2). W tym celu potrzebujemy informacji o preferencjach użytkowników wydobytych z ich wcześniejszych anotacji. Z racji zauważalnego braku spersonalizowanych metod w rozpoznawania poczucia humoru w dziedzinie przetwarzania poczucia humoru, modele omawiane w niniejszym rozdziale zostały opracowane w ramach współpracy w grupie badawczej HumaNLP.

Dokładne wyniki badań eksperymentalnych tych modeli wraz z dokładniejszymi informacjami na ich temat zostały opublikowane w ramach pracy (Kocoń i in., 2021b), gdzie używano nazwy *HuBi - Human Bias*. Dodatkowo, w pracy (Bielaniewicz i in., 2022b) architektury te funkcjonuja pod zmienioną nazwą *Deep-SHEEP*, gdzie skupiono się wyłącznie na zadaniu personalizacji w zadaniu humoru w dziedzinie przetwarzania języka naturalnego. W tym samym zadaniu (personalizacja w klasyfikacji humoru) wykorzystno opisane poniżej modele. Są one dalej użyte w badaniach eksperymentalnych opisanych w dalszych rozdziałach.

5.1 ONEHOT

Model ten (Rys. 6) składa się z najprostszej, a jednocześnie bardzo skutecznej metody reprezentacji anotatorów. Wykorzystuje informacje o identyfikatorze użytkownika w celu włączenia go do wektora jednopunktowego. Prosty wariant architektury apersonalizowanej.



Rysunek 6: Model *OneHot* wykorzystujący zakodowany w postaci one-hot identyfikator użytkownika. (Źródło: (Kazienko i in., 2023)).

5.2 HUBI-FORMULA

Architektura (Rys. 7) wykorzystuje wiedzę o anotatorze u dla zadania humoru t poprzez metrykę interpretowaną jako ludzką charakterystykę *Human Bias*, tj. $HB(u,t)$

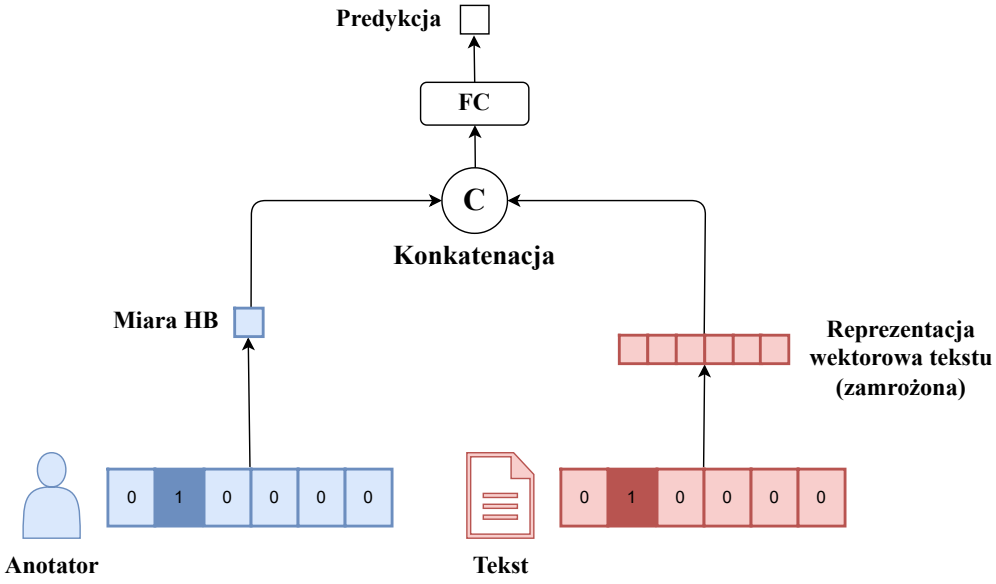
$$HB(u,t) = \frac{\sum_{d \in D_u^{past}} \frac{v_{t,d,u} - \mu_{t,d}}{\sigma_{t,d}}}{|D_u^{past}|} \tag{1}$$

gdzie D_u^{past} jest zbiorem dokumentów $d \in D^{past}$ zaanotowanych przez użytkownika u.

Miara HB dostarcza nam wiedzy na temat unikalnych poglądów i preferencji poszczególnego anotatora. Została ona wykorzystana ściśle w zadaniu humoru jako dodatkowa cecha wejściowa w modelu *HuBi-Formula*.

Wartość ta trafia bezpośrednio na wejście modelu, a dla każdego użytkownika wyliczana jest jednorazowa estymacja, oparta na wskaźniku Z-score. Kiedy HuBi-Formula

łączy się z dowolnym modelem językowym, procesy uczenia się i rozumowania opierają się wyłącznie na informacjach o użytkowniku.



Rysunek 7: Model *HuBi-Formula* wykorzystujący wstępnie obliczoną miarę *HB*. (Źródło: (Kazienko i in., 2023)).

5.3 HUBI-SIMPLE

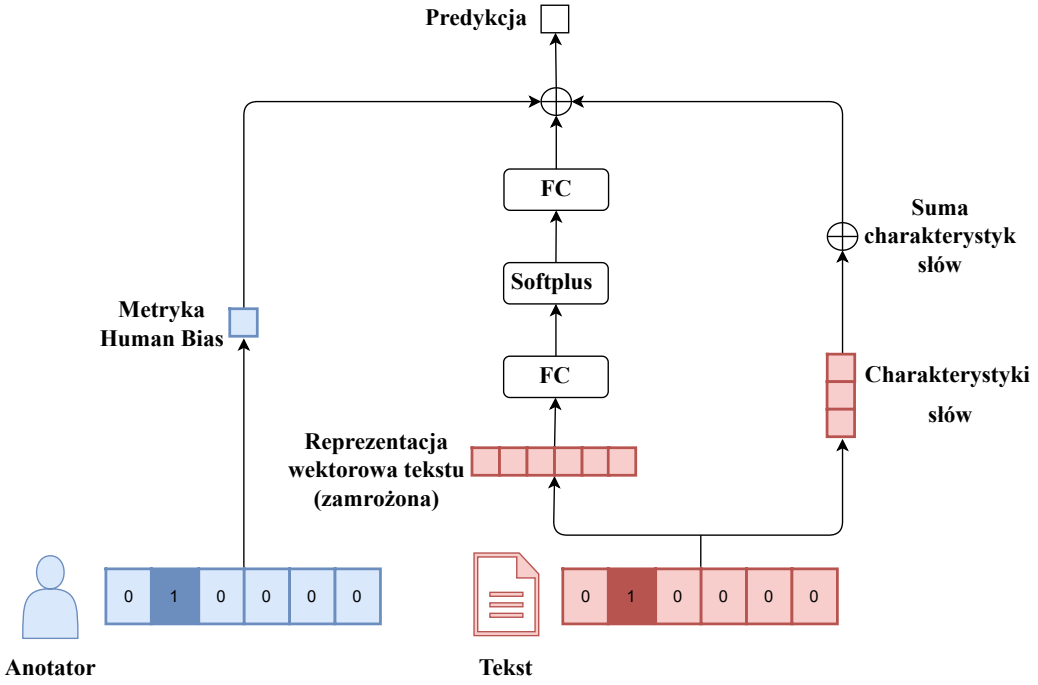
HuBi-Simple (Rys. 8) to prosty, spersonalizowany model, który opiera się na przewidywaniu tekstu na podstawie modelu referecyjnego oraz uwzględnienia preferencji danej osoby i słów użytych w tekście. Definiuje się go w następujący sposób:

$$y(t,u) = a(W_T x_t) + b_u + \sum_{\text{słowo} \in t} b_{\text{słowo}}$$

gdzie:

- t i u oznaczają oceniany tekst i anotatora,
- b_u i $b_{\text{słowo}}$ to wektory charakterystyk związane z danym zadaniem dla anotatora u i słów,
- W_T i x_t oznaczają wektor wag warstwy neuronów oraz reprezentację wektorową ocenianego tekstu t,
- a to funkcja aktywacji *SoftPlus*.

Błędy systematyczne są uwzględniane w ostatecznej predykcji, co można zinterpretować jako dostosowanie wnioskowania dla konkretnej osoby poprzez uwzględnienie jej średniej anotacji w zbiorze danych. Odchylenia są ustawiane losowo na początku, a następnie uczą się poprzez propagację wsteczną. Model ten nie skupia się na relacji między użytkownikiem a zadanym tekstem ani na samym tekście, a jedynie na wyuczonych odpowiednikach statystykach ludzkich anotacji.



Rysunek 8: *HuBi-Simple*: wyuczony model charakterystyk anotatorów. (Źródło: (Kazienko i in., 2023)).

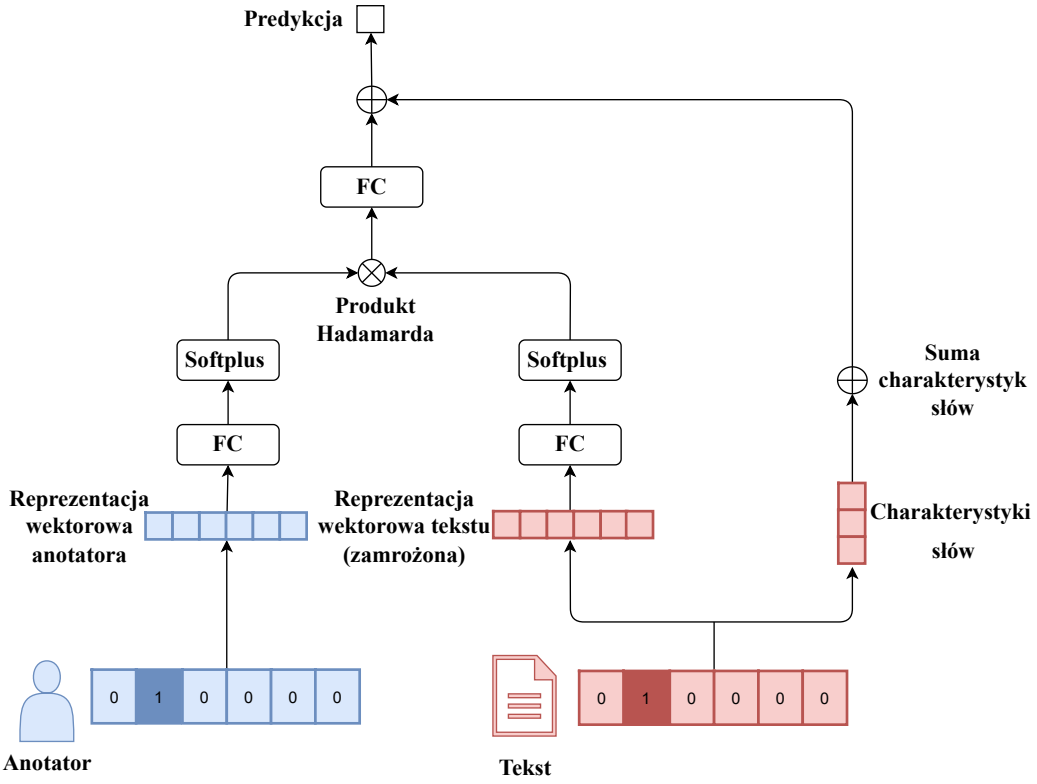
5.4 HUBI-MEDIUM

Ten model (Rys. 9) wyucza się wielowymiarowej, ukrytej reprezentacji wektorowej anotatora, aby uchwycić wiedzę na temat jego poczucia humoru. Aby rozwiązać problem wydajności personalizacji podczas przetwarzania nieznanych tekstów podczas uczenia modelu, HuBi-Medium wykorzystuje reprezentację wektorową człowieka. Pozwala to uchwycić ich podejście do konkretnego tekstu w kontekście poczucia humoru. Predykcja modelu HuBi-Medium jest zdefiniowana w następujący sposób:

$$y(t,u) = W_{TU}(a(W_T x_t) \otimes a(W_U x_u)) + \sum_{word \in t} b_{word}$$

gdzie:

- x_u oznacza wektorową reprezentację danego użytkownika u,
- W_{TU}, W_T, W_U są wagami warstw w pełni połączonych (FC),
- a jest funkcją aktywacji *SoftPlus*,
- $\otimes$ oznacza iloczyn Hadamarda.



Rysunek 9: *HuBi-Medium*: wyuczony model reprezentacji wektorowych anotatorów. (Źródło: (Kazienko i in., 2023)).

6 NOWE METODY UWZGLĘDNIAJĄCE SPERSONALIZOWANE POCZUCIE HUMORU

W tym rozdziale przyjrzymy się nowym metodom, które zostały opracowane, aby efektywniej uwzględniać indywidualne preferencje i unikalne cechy spersonalizowanego humoru. Te autorskie metody proponują innowacyjne podejścia do analizy i klasyfikacji treści humorystycznych w kontekście różnorodnych perspektyw użytkowników.

Pierwsza z metod, Metoda GRU, przetwarza sekwencje tekstowe i informacje o ocenach anotatorów, wykorzystując trzy warstwy GRU do analizy humoru. Pierwsza warstwa koncentruje się na wzorcach tekstowych, druga na ocenach anotatorów, a trzecia łączy te informacje dla zaawansowanej klasyfikacji. Regularizacja dropout i funkcja aktywacji Softplus wspierają generalizację i uchwycenie subtelnych wzorców humorystycznych. Ostateczna klasyfikacja jest dokonywana przez warstwę w pełni połączoną.

Druga z metod, Architektura oparta o Wielogłowicową Atencję, wykorzystuje mechanizm atencji do analizy różnych części tekstu i uwzględnienia różnorodnych perspektyw anotatorów. Model ten posiada cztery głowice atencji, które równocześnie analizują różne cechy humorystyczne. Tekst jest przetwarzany przez warstwę liniową i funkcję aktywacji Softplus, a reprezentacje anotatorów przez dwie warstwy liniowe z regularizacją dropout. Końcowa warstwa liniowa generuje prognozy dotyczące humorystyczności i klasyfikuje typy humoru.

Obie metody przyczyniają się do rozwoju spersonalizowanych systemów analizy humoru, umożliwiając lepsze dostosowanie treści do indywidualnych gustów i oczekiwań użytkowników. Ich innowacyjne podejście do integracji danych wejściowych oraz uwzględnienia różnorodnych perspektyw oferują nowe możliwości w tworzeniu bardziej zaawansowanych i efektywnych narzędzi do analizy humoru w kontekście uczenia maszynowego.

6.1 ARCHITEKTURA GRU

Model GRU (ang. *Gated Recurrent Unit*) (Rys. 10) jest zaprojektowany w celu analizy i oceny treści pod kątem charakterystyki humorystycznej badanych treści. Jego celem jest przetwarzanie sekwencyjnych danych, łącząc informacje z różnych źródeł (tekst i anotacje) oraz ucząc się zaawansowanych wzorców poprzez zastosowanie warstw GRU i technik regularizacyjnych. Dzięki zastosowaniu kilku warstw GRU i mechani-

zmów takich jak dropout oraz funkcja aktywacji, model jest w stanie uchwycić złożoność i niuanse w danych wejściowych. Metodę GRU można opisać w następujący sposób:

$$y(t,u) = W_{TU}(a(GRU_{U2}(GRU_{U1}(x_u))) \otimes a(GRU_T(x_t)))$$

gdzie:

- t i u oznaczają oceniany tekst i anotatora,
- x_u oznacza reprezentację wektorową danego użytkownika u,
- GRU_{U1} oznacza wektor wag komórki GRU otrzymującej na wejściu reprezentację wektorową użytkownika x_u,
- GRU_{U2} oznacza wektor wag komórki GRU otrzymującej na wejściu wyjście z komórki GRU GRU_{U1},
- x_t oznacza reprezentację wektorową ocenianego tekstu t,
- GRU_T oznacza macierz wag komórki GRU otrzymującej na wejściu reprezentację wektorową tekstu x_t,
- a to funkcja aktywacji *SoftPlus*,
- W_{TU} oznacza wektor wag warstwy neuronów otrzymującej na wyjściu wynik iloczynu Hadamarda obliczonego dla wartości funkcji aktywacji a dla wyjścia z komórki GRU_{U2} i dla wyjścia z komórki GRU_T.

Na początku model zaczyna od przetworzenia identyfikatorów anotatorów (czyli osób oceniających treść) w reprezentacje wektorowe. Każdy anotator ma swoją unikalną reprezentację, co pozwala modelowi uwzględnić różne perspektywy lub oceny humoru. Taki mechanizm umozliwia modelowi rozroznienie wiedzy na temat różnych anotatorow.

Następnie przechodzimy przez wszystkie warstwy GRU.

Pierwsza Warstwa GRU analizuje tekstowe osadzenia treści humorystycznych. GRU przetwarza sekwencje słów w treci, identyfikując wzorce i struktury, które mogą sugerować elementy humorystyczne, takie jak ironia, sarkazm czy gry słowne.

Druga Warstwa GRU skupia się na wektorowych reprezentacjach anotatorów, przetwarzając informacje o oceniających. Ta warstwa może dostarczyć kontekst dla indywidualnych ocen lub preferencji w ocenie humoru.

Trzecia Warstwa GRU integruje i przetwarza informacje z drugiej warstwy GRU. Umożliwia modelowi łączenie wiedzy o treści z różnymi ocenami anotatorów, co pozwala na bardziej zaawansowaną analizę i klasyfikację treści humorystycznych.

Następnie wykorzystywana jest Regularyzacja *Dropout*, aby uniknąć nadmiernego dopasowania modelu do danych treningowych, co mogłoby prowadzić do przeuczenia. W kontekście klasyfikacji humoru, regularizacja ta pomaga modelowi generalizować swoje umiejętności na nowe, nieznane treści humorystyczne.

W kolejnym kroku funkcja aktywacji *Softplus* wprowadza nieliniowość do modelu, co pozwala na uchwycenie bardziej złożonych i subtelnych wzorców w danych tekstowych. Jest to istotne w klasyfikacji humoru, gdzie nieliniowe i złożone relacje między słowami mogą wskazywać na humorystyczne treści.

Na końcu wyniki z pierwszej i trzeciej warstwy GRU są łączone i przekazywane przez warstwę w pełni połączoną *(Fully Connected)*. To pozwala na uzyskanie ostatecznej klasyfikacji treści. Wyjściowa warstwa modelu może dostarczać prognozy na temat tego, czy dana treść jest humorystyczna, czy nie, lub jakie są jej cechy humorystyczne.

Architektura ta, w kontekście klasyfikacji treści humorystycznych, wykorzystuje zaawansowane techniki przetwarzania sekwencyjnego i osadzenia, aby skutecznie analizować i klasyfikować teksty pod kątem humoru. Integrując informacje z różnych warstw GRU i uwzględniając różne perspektywy anotatorów, model ten jest w stanie uchwycić subtelności i niuanse, które definiują humorystyczną treść. Regularizacja dropout i funkcje aktywacji pomagają w generalizacji i uchwyceniu złożonych wzorców, co jest kluczowe w skutecznej klasyfikacji humoru.

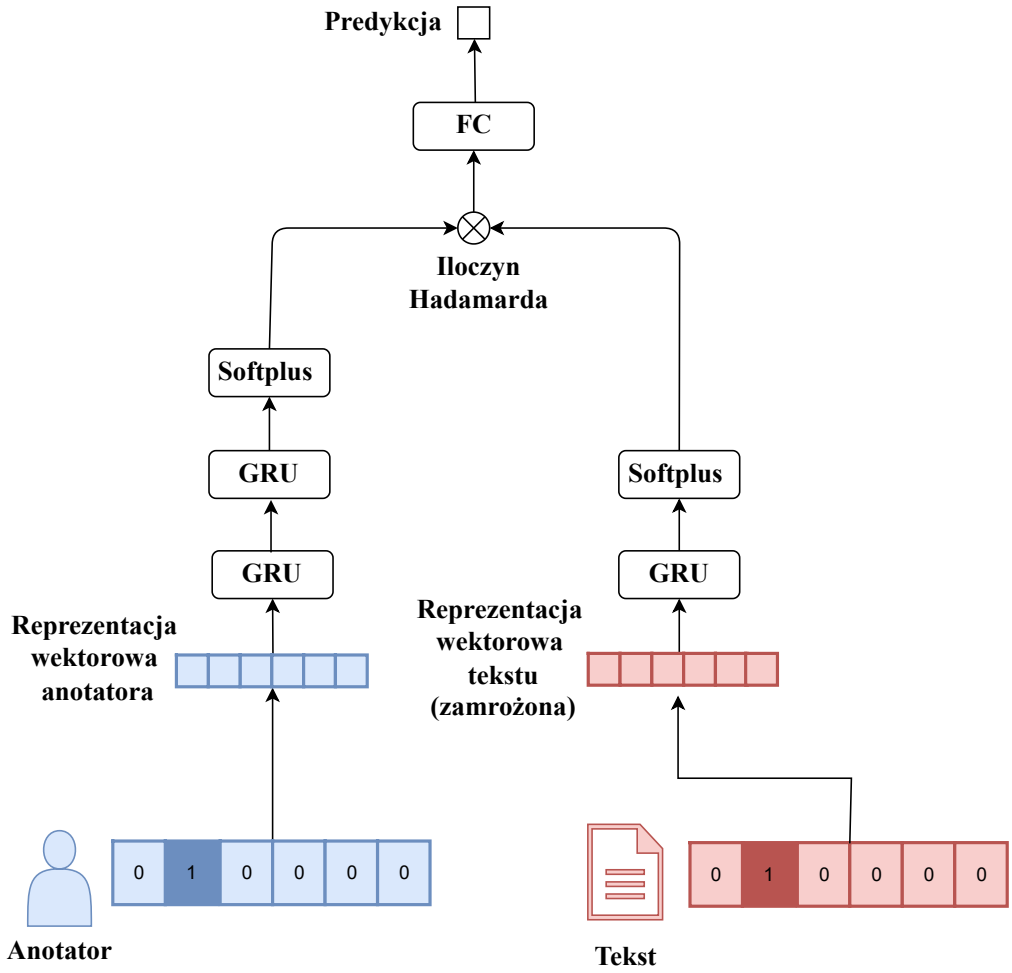
6.2 ARCHITEKTURA BAZUJĄCA NA WIELOGŁOWICOWEJ ATENCJI

Architektura z Wielogłowicową Atencję (ang. *Multi-head attention*) (Rys. 11) jest zaprojektowana do klasyfikacji treści humorystycznych. Wykorzystuje zaawansowane mechanizmy przetwarzania sekwencyjnego i atencji, aby uchwycić subtelności i złożoność humoru w treściach tekstowych. Integruje różne perspektywy anotatorów i analizuje różne aspekty treści za pomocą wielogłowicowego mechanizmu atencji.

Mechanizm atencji pozwala modelowi skoncentrować się na różnych częściach sekwencji wejściowej podczas przetwarzania informacji. W kontekście modelowania sekwencyjnego, atencja pomaga modelowi zrozumieć, które elementy sekwencji są istotne w danym kontekście.

Wiele głowic (ang. *Multi-head*) działa na podstawie rozdzielenia mechanizmu atencji na kilka "głowic", z których każda uczy się innej charakterystyki relacji między elementami sekwencji. W zaprojektowanej architekturze model ma cztery takie głowice atencji, które równocześnie analizują dane wejściowe, a następnie ich wyniki są łączone. Architekturę modelu można opisać następującym wzorem:

$$y(t,u) = W_{TU}(MHA(a(W_U(x_u)) \frown a(W_T(x_t))))$$



Rysunek 10: Model *GRU* integrujacy pojedynczą warstwę GRU reprezentacji wektorowej tekstu i podwójną warstwę GRU reprezentacji wektorowej anotatora. (Źródło: Opracowanie własne).

gdzie:

- t i u oznaczają oceniany tekst i anotatora,
- x_u oznacza reprezentację wektorową danego użytkownika u,
- W_U oznacza wektor wag warstwy neuronów otrzymującej na wejściu reprezentację wektorową użytkownika x_u,
- x_t oznacza reprezentację wektorową ocenianego tekstu t,
- W_T oznacza wektor wag warstwy neuronów otrzymującej na wejściu reprezentację wektorową tekstu x_t,
- a to funkcja aktywacji *SoftPlus*,

- $\frown$ oznacza operację konkatenacji dwóch wektorów,
- MHA oznacza macierz wag komórki atencji wielogłowicowej zawierającej 4 głowice atencji,
- W_{TU} oznacza wektor wag warstwy neuronów otrzymującej na wyjściu wyjście z komórki atencji wielogłowicowej MHA.

Działanie modelu rozpoczyna się od przetwarzania identyfikatorów anotatorów w gęste wektory. Anotatorzy, którzy oceniają treści humorystyczne, mogą mieć różne opinie i preferencje. Dlatego reprezentowanie tych różnic w przestrzeni wektorowej pozwala modelowi uwzględnić te zróżnicowane perspektywy w analizie treści humorystycznych.

Następnie przechodzimy do przetwarzania wejściowego, które składa się z kilku kroków. Pierwszy polega na tym, że tekstowe dane wejściowe, które mogą zawierać żarty, gry słowne, sarkazm i inne formy humoru, są najpierw przetwarzane przez warstwę liniową. Ta warstwa zwiększa wymiar danych, umożliwiając modelowi uchwycenie bardziej złożonych wzorców. Funkcja aktywacji *Softplus* dodaje nieliniowość, co jest ważne dla uchwycenia subtelnych niuansów humoru.

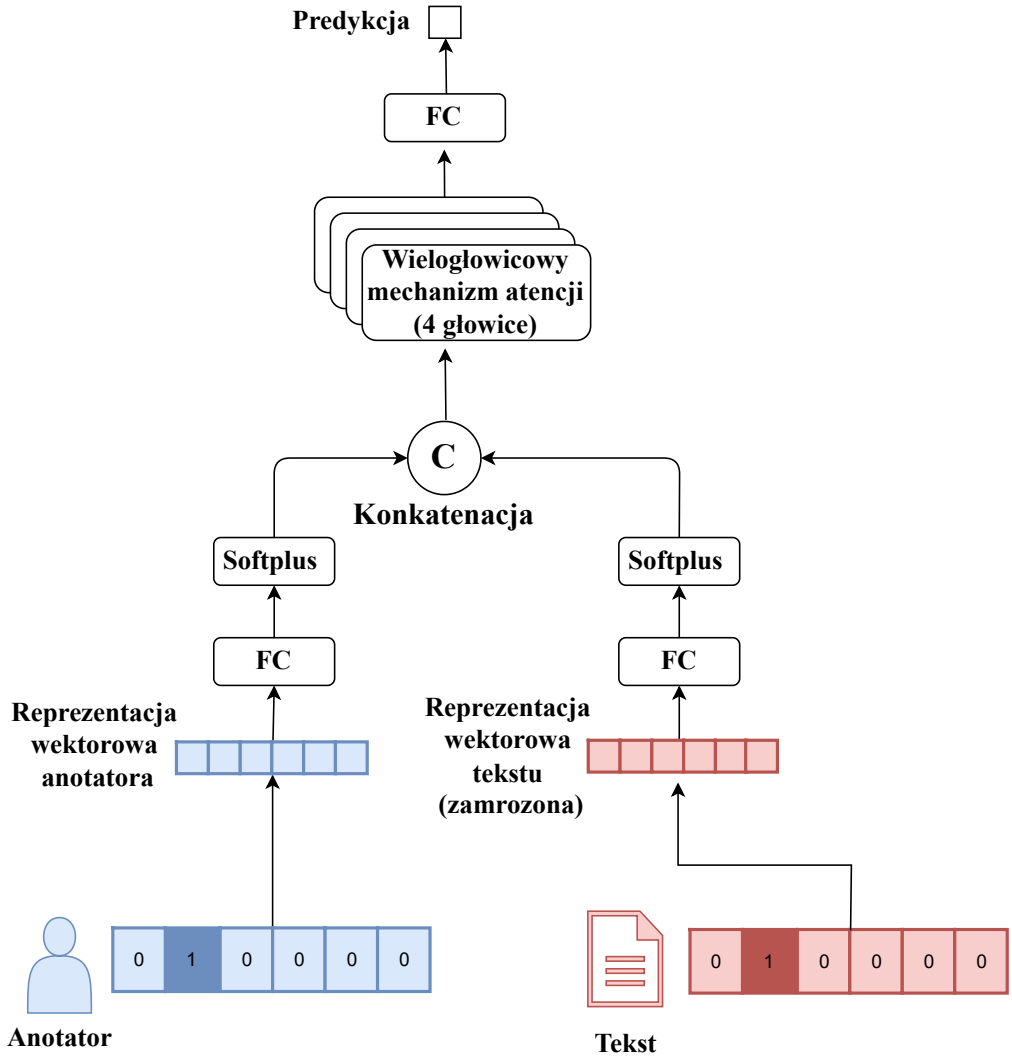
Natomiast reprezentacje wektorowe anotatorów są przetwarzane przez dwie warstwy liniowe. Regularizacja dropout jest stosowana, aby uniknąć nadmiernego dopasowania modelu do specyficznych opinii anotatorów, co pozwala na lepszą generalizację. Następnie, te osadzenia są przetwarzane w celu uzyskania końcowych reprezentacji, które mogą być używane w analizie treści.

Następnie przechodzimy do mechanizmu atencji (ang. *Attention*), w którym następuje przygotowanie Kluczy, Kwerend i Wartości (ang. *Key, Query, Value*). Dla każdego elementu treści humorystycznej przygotowywane są kwerendy, klucze i wartości. W kontekście humoru, mechanizm ten umożliwia modelowi uwzględnienie różnych aspektów treści oraz różnorodnych ocen anotatorów w analizie humoru.

Wielogłowicowy mechanizm atencji (ang. *Multi-head Attention*) z czterema głowami pozwala modelowi uchwycić różne aspekty i typy humoru w treści. Dzięki temu model może jednocześnie analizować różne cechy humorystyczne, takie jak ironia, gry słowne, czy sarkazm, poprawiając dokładność klasyfikacji.

Następnie po przetworzeniu treści przez mechanizm atencji, wyniki są przekazywane przez końcową warstwę liniową, która mapuje je na wymiar wyjściowy modelu. Ta warstwa generuje końcowe prognozy, które określają, czy dana treść jest humorystyczna, oraz może klasyfikować różne typy humoru.

Dodatkowo obliczane są skojarzenia między kwerendami a kluczami. Wyniki tych obliczeń mogą być użyte do analizy podobieństw między różnymi segmentami treści humorystycznej, co może wspierać dokładniejsze klasyfikowanie i zrozumienie humoru.



Rysunek 11: Model *Wielogłowicowego mechanizmu atencji* zrównolegla proces analizy, a następnie scala uchwycone charakterystyki. (Źródło: Opracowanie własne).

Część III

BADANIA

7 BADANIA EKSPERYMENTALNE

Przeprowadzone badania opierały się na analizie i ocenie skuteczności różnych technik przetwarzania języka naturalnego (NLP) w zadaniu rozpoznawania i modelowania humoru w tekstach. Głównym celem było zbadanie, jak różne podejścia, takie jak porównanie predykcji modeli językowych oraz modeli predykcyjnych, transfer uczenia, analiza wpływu rozmiaru zbioru uczącego, porównanie modeli dedykowanych do modelu generatywnego ogólnego przeznaczenia mogą być wykorzystane do lepszego zrozumienia i przewidywania humorystycznych treści.

W ramach tych badań przeprowadzono szereg eksperymentów, które miały na celu ocenę, w jakim stopniu te techniki mogą pomóc w tworzeniu modeli, które skutecznie rozpoznają i kwantyfikują humor. Kluczowym elementem było także zbadanie, jak różne zbiory danych, w tym te zebrane z różnych źródeł i oznaczone przez użytkowników, mogą być integrowane w celu poprawy jakości predykcji.

7.1 PLAN BADAŃ

W ramach planu badań opracowano trzy kluczowe scenariusze eksperymentalne, które mają na celu kompleksową analizę modelowania śmieszności w zadaniach przetwarzania języka naturalnego (NLP). Każdy z tych scenariuszy skupia się na innym aspekcie uczenia maszynowego w kontekście zadania śmieszności, umożliwiając zbadanie różnych technik i ich wpływu na jakość predykcji humoru w tekstach.

W pierwszym scenariuszu badamy jak efektywne są nowe architektury spersonalizowane (GRU oraz Attention, Podrozdział 6) względem innym architektur spersonalizowanych (Rozdział 5) dla różnych modeli językowych reprezentacji tekstu i dla zbiorów o różnej charakterystyce, w tym tekstów w języku zarówno anielskim jaki i polskim (Podrozdziały 7.3.1 & 7.5.1). Zbiory polskie Doccano 1 (Podrozdział 7.2.1.6) oraz Doccano 2 (Podrozdział 7.2.1.7) były specjalnie zgromadzone w ramach projektu CLARIN przez grupę HumaNLP, której jestem członkiem. Wszystkie spersonalizowane architektury są porówane z Modelem Referencyjnym – zgeneralizowanym, który nie posiada żadnej informacji o użytkowniku (Podrozdział 4.2).

Drugi scenariusz ma na celu zbadanie wpływu rozmiaru zbioru uczącego na jakość predykcji. W tym kontekście analizowane jest, jak różne wielkości zbiorów treningowych wpływają na efektywność modeli w zadaniach klasyfikacji i regresji. Badanie to jest istotne, ponieważ w rzeczywistych zastosowaniach NLP często mamy do czynienia z ograniczonymi danymi. W związku z tym, zrozumienie, jak zwiększenie lub

zmniejszenie ilości danych wpływa na zdolność modeli do rozpoznawania humoru, może dostarczyć cennych informacji na temat optymalizacji procesu treningu.

Trzeci scenariusz dotyczy transferu uczenia i badania wpływu technik transferu uczenia na jakość predykcji śmieszności. Transfer uczenia polega na wykorzystaniu wiedzy zdobytej przez model na jednym zadaniu lub w jednym zestawie danych, aby poprawić jego wydajność na innym, często pokrewnym zadaniu. W kontekście tego badania, transfer uczenia może obejmować przenoszenie wiedzy z modeli trenowanych na dużych, ogólnych zbiorach danych do bardziej specyficznych zadań związanych z humorem. Scenariusz ten pozwala na ocenę, czy i w jakim stopniu techniki transferu uczenia mogą poprawić zdolność modeli do efektywnej predykcji śmieszności tekstów w różnych kontekstach i językach.

Czwarty scenariusz to testowanie modelu generatywnego ChatGPT-3.5 w uogólnionym zadaniu rozpoznawania humoru dla dwóch zbiorów danych (ColBERT i Sarcasmania). Badania związane z humorem były częścią rozległych prac naukowych przeprowadzonych w ramach projektu CLARIN, których efektem była pierwsza na świecie wielkoskalowa analiza ChatGPT-3.5 poprzez 25 zadań i łącznie ponad 48 tys., promtów, w tym wspomnianych dwóch zadań dotyczących śmieszności, za które ja odpowiadałam (Kocoń i in., 2023). Prace były przeprowadzone w styczniu i lutym 2023 roku.

Każdy z tych scenariuszy dostarcza innej perspektywy na problem modelowania humoru w przetwarzaniu języka naturalnego, pozwalając na pełniejsze zrozumienie, jak różne techniki i podejścia mogą być wykorzystane do automatycznej analizy i rozpoznawania humorystycznych treści.

7.2 ZBIORY DANYCH

W celu zgromadzenia odpowiednio dużej liczby danych potrzebnych do kompleksowej analizy poczucia humoru, zebrano publicznie dostepne źródła danych, a także przeprowadzono procedury anotacyjne

Wybrano 9 zbiorów danych zawierających anotowane fragmenty tekstu humorystycznego w postaci pojedynczych słów, fraz, a nawet par tekstów przed i po drobnej zmianie. Pięć z tych zestawów danych było ściśle spersonalizowanych, co oznacza, że zawierało informacje o identyfikatorach użytkowników oraz licznych anotacjach dla niezależnych tekstów, Podrozdział 7.2.1. Są to główne zbiory wykorzystane do testowania jakości wnioskowania przez moje i inne architektury spersonalizowane. Dwa z pięciu zbiorów są w języku polskim i zostały przeze mnie zgromadzone we współpracy z grupa HumaNLP.

Pozostałe cztery zestawy danych są uogólnione, z dostępnością tylko jednej anotacji na tekst, Podrozdział 7.2.2. Zdecydowana większość tekstów była w języku an-

gielskim, ale występowały także teksty w języku hiszpańskim oraz polskim. Bardziej szczegółowe informacje na temat zbiorów danych spersonalizowanych znajdują się w Tab. 1 & Tab. 2, a informacje o zbiorach zgeneralizowanych znajdują się w Tab. 3 oraz Tab. 4.

Zbiory zgeneralizowane były wykorzystane w badaniach transferu uczenia w celu sprawdzenia jak wiele wiedzy na temat charakterystyki humoru może być uzupełnione przez dane spersonalizowane w 4 różnych konfiguracjach eksperymentalnych (Podrozdziały 7.6.3 oraz 7.7.3).

7.2.1 Zbiory danych spersonalizowanych

Zbiory spersonalizowane charakteryzują się tym, że prócz samego tekstu i anotacji posiadają również odpowiedni identyfikator lub inną formę rozpoznania anotatora, który ową anotację dokonał. Oznacza to zwykle to, że pojdynczy tekst był anotowany przez różne osoby z różną wartością anotacji, tj. różnym odbiorem humoru związanego z danym tekstem.

Zbiory w języku polskim Doccano 1 (Podrozdział 7.2.1.6) oraz Doccano 2 (Podrozdział 7.2.1.7) zostały przeze mnie zgromadzone w ramach projektu CLARIN we współpracy z grupą HumaNLP. Anotatorami byli przeszkoleni, ale zwykli pracownicy administracyjni Politechniki Wrocławskiej. Proces zebrania danych wymagał opracowania scenariuszy anotacji oraz zorganizowania środowiska dla anotatorów. Analizy obu pozyskanych zbiorów zostały opublikowane w pracy (Bielaniewicz i Kazienko, 2023). Zbiór Doccano 1 był dodatkowo wykorzystany w naszej pracy (Mieleszczenko-Kowszewicz i in., 2023).

7.2.1.2 Cockamamie Gobbledegook

Zbiór Cockamamie Gobbledegook zaprezentowany w (Engelthaler i in., 2018) zawiera 10 tysięcy angielskich wyrażeń (1-2 wyrazy) opatrzonych anotacjami przez amerykańskich pracowników Amazon Mechanical Turk. Każdy anotator ocenił dane słowo jako *nieśmieszne* (0) lub *śmieszne* (1). Na potrzeby badań wykorzystano tylko te słowa, które miały co najmniej jedną pozytywną i jedną negatywną adnotację, aby lepiej uchwycić różnice w preferencjach użytkowników. Informacje na temat zbioru znajdują się w Tab. 1.

7.2.1.3 HUMOR

Zbiór HUMOR składa się z 27 tysięcy hiszpańskich tweetów oznaczonych przez pracowników crowd-sourcingu jako *śmieszne* (1-5) lub *nieśmieszne* (0) (Castro i in., 2018). Jeśli tweet został uznany przez osobę za humorystyczny, musiała ona dodat-

kowo określić, jak bardzo jest śmieszny, w skali od 1 (mało zabawne) do 5 (bardzo śmieszne). Na potrzeby eksperymentów użyto wersji[1] zbioru danych zawierającej teksty, a nie tylko identyfikatory tweetów. Ponadto wzięto pod uwagę tylko te teksty, które miały więcej niż jedną anotację. Ostateczny zbiór danych użyty w badaniach składał się z 26 967 anotacji na 8 284 tekstach wykonanych przez 3 137 osób. Informacje na temat zbioru znajdują się w Tab. 1.

7.2.1.4 Jester

Zbiór Jester jest zazwyczaj wykorzystywany w zadaniach rekomendowania dowcipów (Chakrabarty i in., 2019). Charakter danych, który dokładnie identyfikuje anotatorów i ich oceny, pozwala testować metody personalizacji w wykrywaniu humoru. Zawiera on jedynie 100 tekstowych dowcipów, ocenionych przez ponad 73 tysiące użytkowników poprzez wartości z zakresu [-10,10]. Całkowita liczba anotacji to 4,1 miliona ocen treści tekstowych (Goldberg i in., 2001). Informacje na temat zbioru znajdują się w Tab. 2.

7.2.1.5 Humicroedit

Zbiór Humicroedit został opublikowany w ramach konkursu SemEval-2020 na rozpoznawanie humoru w ramach zadania nr. 7 (Hossain i in., 2020). Autorzy wykorzystali anotatorów z Amazon Mechanical Turk do modyfikacji pojedynczego słowa w 14 tysiącach nagłówków wiadomości, aby zmienić je tak, by wywoływały rozbawienie. Następnie pięciu innych anotatorów oceniało każdy edytowany nagłówek w skali od 0 do 3, gdzie 0 oznaczało brak rozbawienia, a wartość 3 oznaczała bardzo śmieszną treść. W związku z tym, osadzenia zarówno oryginalnych, jak i zmodyfikowanych nagłówków wraz z ich ocenami są dostarczane na wejściu do wszystkich modeli, aby uwzględnić kontekst w zadaniu rozpoznawania humoru (Hossain i in., 2019). Informacje na temat zbioru znajdują się w Tab. 1.

7.2.1.6 Doccano 1

Zbiór danych Doccano 1 jest jedną z wersji projektu Doccano 1.0, który bada postrzeganie i emocje związane z treściami tekstowymi. Każdy tekst w tym zbiorze ma długość do 132 słów (μ = 24,5, σ = 16,2). Średnio jedna osoba oznaczała około 790 tekstów, a przeciętnie każdy tekst był oceniany przez około 32 osoby. Łącznie zebrano około 31 700 anotacji, z których każda obejmuje 26 różnych wymiarów: (1) *spokój*, (2) *współczucie*, (3) *zadowolenie*, (4) *inspiracja*, (5) *radość*, (6) *pozytywność*, (7) *zdziwienie*, (8) *złość*, (9) *obrzydzenie*, (10) *strach*, (11) *negatywność*, (12) *smutek*, (13) *zgoda*, (14) *zakłopota-*

1 https://github.com/pln-fing-udelar/humor/tree/main/previous

nie, **(15) *śmieszne dla mnie***, (16) *śmieszne dla kogoś*, (17) *niezrozumiałe*, (18) *interesujące*, (19) *ironia*, (20) *obraźliwe dla mnie*, (21) *obraźliwe dla kogoś*, (22) *polityczne*, (23) *sympatia*, (24) *zrozumiałe*, (25) *zaufanie*, i (26) *wulgarne*. Na potrzeby badań zadania śmieszności skupiono się wyłącznie na wymiarze (15). Jeśli anotator nie miał zdania na temat danego wymiaru, mógł pominąć wybór. Średnio etykiety z wartością zerową pojawiają się w 62% przypadków, z odchyleniem standardowym równym 22%. Puste etykiety występują w 4% przypadków, z odchyleniem standardowym wynoszącym 8%. W ramach wstępnego przetwarzania ze zbioru odfiltrowano teksty, dla których wszyscy anotatorzy wstrzymali się od głosu w przypadku wymiaru **(15) *śmieszne dla mnie***. Informacje na temat zbioru znajdują się w Tab. 2.

7.2.1.7 Doccano 2

Zbiór danych Doccano 2, podobnie jak Doccano 1, jest częścią projektu Doccano 1.0. Zawiera anotacje od 40 osób. Średnio każda osoba oznaczała około 358 tekstów, przy czym każdy tekst był oceniany przez 2 anotatorów. W sumie zebrano około 17 700 anotacji. Tak jak w przypadku Doccano 1, każda anotacja obejmuje 26 niezależnych wymiarów pokrywających się z tym zbiorem. Dodatkowo, podobnie jak w przypadku zbioru Doccano 1, na potrzeby badań zadania śmieszności skupiono się na wymiarze **(15) *śmieszne dla mnie***. Osoby, które nie miały zdania na temat danego wymiaru, miały możliwość pominięcia wyboru. Wartości anotacji były ustalane podczas procedury anotacji w sposób analogiczny do tego, jak miało to miejsce w Doccano 1. W ramach wstępnego przetwarzania ze zbioru odfiltrowano teksty, dla których wszyscy anotatorzy wstrzymali się od głosu w przypadku wymiaru **(15) *śmieszne dla mnie***. Informacje na temat zbioru znajdują się w Tab. 2.

7.2.2 Zbiory danych uogólnionych

Informacje zawarte w uogólnionych zbiorach danych często posiadają nikłą lub brak informacji na temat anotatora. Głównymi punktami uwagi jest tutaj tekst oraz ocena tego tekstu.

7.2.2.1 ColBERT

Zbiór ColBERT to duży zbiór danych utworzony do zadania wykrywania humoru. Wiele istniejących zbiorów danych dotyczących wykrywania humoru łączyło formalne teksty niehumorystyczne z krótkimi, potocznymi tekstami humorystycznymi (Annamoradnejad i Zoghi, 2020). Zbiór ten zawiera 200 000 krótkich tekstów (100 000 pozytywnych i 100 000 negatywnych). Na potrzeby badań z wykorzystaniem modelu generatywnego ChatGPT3.5 wylosowano 1000 tekstów. Po odfiltrowaniu tekstów jed-

noznacznie ocenianych jako nieśmieszne, ze wstępnie wylosowanych tekstów pozostało 995. Informacje na temat zbioru znajdują się w Tab. 3.

7.2.2.2 HahaIberlef2021

Zbiór HahaIberlef2021 to korpus tweetów oznaczonych przez GRUpę użytkowników, podzielony na trzy zestawy: treningowy (24 000 tweetów), deweloperski (6 000 tweetów) i testowy (6 000 tweetów)(Chiruzzo i in., 2021). Anotacja wykorzystuje schemat głosowania, w którym użytkownicy mogą wybrać jedną z sześciu opcji (Chiruzzo i in., 2020): tweet jest niehumorystyczny, lub tweet jest śmieszny, a ocena jest przyznawana w skali od jeden (mało zabawny) do pięć (bardzo śmieszny).Informacje na temat zbioru znajdują się w Tab. 3.

7.2.2.3 Hahackathon

Zbiór danych Hahackathon zawiera 10 000 tekstów oznaczonych pod względem humoru i obraźliwości przez 20 anotatorów w wieku od 18 do 70 lat (Meaney i in., 2021). 80% jego danych pochodzi z Twittera. Pozostałe 20% tekstów zostało wybrane z zestawu danych Kaggle Short Jokes.Informacje na temat zbioru znajdują się w Tab. 4.

7.2.2.4 Sarcasmania

Sarcasmania to zbiór danych pozyskany z Kaggle, składający się z 39 780 tekstów z platformy Twitter (Siddiqui, 2019). Każdy tekst jest oznaczony pod trzema wymiarami: śarkazm", "humor"i óbraźliwość". Aby uwzględnić ten zbiór danych, skoncentrowaliśmy się wyłącznie na aspekcie humorystycznym anotacji w tym zbiorze tekstów. Na potrzeby badań z wykorzystaniem modelu generatywnego ChatGPT3.5 wylosowano 1000 tekstów. Po odfiltrowaniu tekstów jednoznacznie ocenianych jako nieśmieszne, ze wstępnie wylosowanych tekstów pozostało 990. W tym scenariuszu nazwa zbioru została skrócona do wersji Sarcasm. Informacje na temat zbioru znajdują się w Tab. 4.

Tabela 1: Spersonalizowane profile zbiorów danych. Każdy zestaw danych zawiera określoną liczbę etykiet. Szczegóły i dalsze informacje na ich temat można znaleźć w sekcji 7.2.

Właściwości \ Zbiór danych	Cockamamie Gobbledegook	Humor	Humicroedit
Profil danych tekstowych	1-2 słowa	tweety	pary & nagłówki
Liczba tekstów	10884	8284	14886
Liczba anotacji	40673	26967	74430
Liczba anotatorów	351	3137	5
Średnia liczba anotacji na tekst	3.74	3.26	5
Średnia liczba anotatorów na tekst	115.88	57.44	14886
Zbalansowanie klas (0/1)	382083 / 51197	12091 / 9978	31863 / 42572
Język	Angielski	Hiszpański	Angielski

Tabela 2: Spersonalizowane profile zbiorów danych. Każdy zestaw danych zawiera określoną liczbę etykiet. Szczegóły i dalsze informacje na ich temat można znaleźć w sekcji 7.2. Wartości umieszczone w nawiasach dla zbiorów Doccano 1 i Doccano 2, dotyczą niepustych anotacji wymiaru ***śmieszne dla mnie***.

Właściwości \ Zbiór danych	Jester	Doccano 1	Doccano 2
Profil danych tekstowych	krótkie żarty	komentarze	komentarze
Liczba tekstów	100	880 (120)	8891 (200)
Liczba anotacji	4136360	31521 (4815)	17533 (9872)
Liczba anotatorów	73421	39 (39)	49 (49)
Średnia liczba anotacji na tekst	41,363.6	35.81 (40.13)	1.97 (49.36)
Średnia liczba anotacji na anotatora	56.34	808.23 (123.46)	357.81 (201.47)
Język	Angielski	Polski	Polski

7.3 SCENARIUSZE BADAŃ

W celu przeprowadzenia badań nad poczuciem humoru opracowano następujące scenariusze eksperymentalne aby zwrócić uwagę na istotne aspekty takie jak personalizacja humoru czy różnica ich jakości predykcji w porównaniu ze zgeneralizowanymi modelami uczenia maszynowego w celu zbadania:

Tabela 3: Profile niespersonalizowanych zbiorów danych. Każdy zestaw danych zawiera określoną liczbę etykiet. Liczba adnotacji jest równa liczbie testów ze względu na brak jakiegokolwiek rozróżnienia użytkownika. Szczegóły i dalsze informacje na ich temat można znaleźć w sekcji 7.2. Liczba tekstów dla zbioru ColBERT, która jest widoczna w nawiasie jest liczbą, jaka została uwzględniona w badaniach dla scenariusza z modelem generatywnym ChatGPT-3.5.

Zbiór danych / Właściwości	ColBERT	HahaIberlef2021
Profil treści tekstowych	krótkie zdania	tweety
Liczba tekstów	199996 (995)	33824
Zbalansowanie klas (0/1)	100000 (500) / 99996 (495)	20778 / 13046
Język	Angielski	Hiszpański

Tabela 4: Uogólnione profile zbiorów danych. Każdy zestaw danych zawiera określoną liczbę etykiet. Liczba adnotacji jest równa liczbie testów ze względu na brak jakiegokolwiek rozróżnienia użytkownika. Szczegóły i dalsze informacje na ich temat można znaleźć w sekcji 7.2. Liczba tekstów dla zbioru Sarcasmania, która jest widoczna w nawiasie jest liczbą, jaka została uwzględniona w badaniach dla scenariusza z modelem generatywnym ChatGPT-3.5.

Zbiór danych / Właściwości	Hahackathon	Sarcasmania
Profil treści tekstowych	tweety & żarty	tweety
Liczba tekstów	7958	39778 (990)
Zbalansowanie klas (0/1)	3068 / 4890	20135 (500) / 19643 (490)
Język	Angielski	Angielski

- Wpływu różnych spersonalizowanych głębokich architektur predykcyjnych i modeli językowych na jakość predykcji śmieszności,
- Wpływu rozmiaru zbioru uczącego na jakość predykcji śmieszności,
- Wpływu transferu uczenia na jakość predykcji śmieszności.
- Analiza jakości zgeneralizowanej klasyfikacji treści humorystycznych przez ogólny model generatywny ChatGPT-3.5.

.

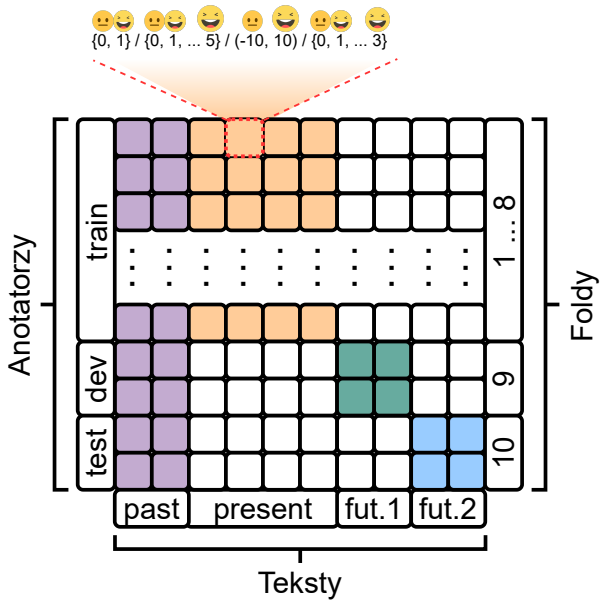
Dla każdego scenariusza, eksperymenty przeprowadzono w dwóch konfiguracjach: (1) klasyfikacja binarna z miarami Macro F1 i F1 dla pozytywnej klasy śmieszności oraz (2) regresja z miarą MSE (średni błąd kwadratowy). Aby zmierzyć procentowy wzrost wyniku F1, zastosowano $\mathrm{F1}roznica = (\mathrm{F1}model - \mathrm{F1}baseline) \cdot 100$. W przypadku regresji wykorzystano zysk średniego błędu kwadratowego obliczony za pomocą $\mathrm{MSE}roznica = (\mathrm{MSE}model - \mathrm{MSE}baseline) \cdot 100$. W przedstawionych wzorach *baseline* dla miar F1 i MSE oznacza wynik uzyskany dla modelu referencyjnego dla zadań zarówno klasyfikacji, jak i regresji. Każdy potrójny zestaw <*zbiór danych, model językowy, model personalizacyjny*> został oceniony przy użyciu walidacji krzyżowej na 10 zbiorach użytkowników.

Zróżnicowana wielkość oraz charakter anotowanych tekstów (słowotwórstwo w zbiorze Cockamamie, krótkie żarty w zbiorze Jester) pozwoliły zweryfikować, czy rozwiązania dobrze sprawdzają się zarówno w pracy nad pojedynczymi słowami, jak i pełnymi tekstami. Aby uniknąć trenowania na znanych anotatorach i tekstach oraz w celu uzyskania lepszej uogólnialności modelu testowego, każdy zbiór danych podzielono na 10 rozłącznych części do walidacji krzyżowej w oparciu o użytkowników, a także na niezależne, rozłączne podzbiory tekstów (Rysunek 12).

W przypadku badań dla ChatGPT, jeśli to możliwe, próbowano walidować ChatGPT przy użyciu jednej miary — F1 Macro, która jest powszechnie akceptowana dla danych niezbalansowanych. F1 Macro w klasyfikacji wieloetykietowej jest średnią harmoniczną między precyzją a czułością obliczaną dla każdej etykiety. Jeśli Q to liczba etykiet, p_i oraz r_i to precyzja i czułość obliczane dla etykiety i, F1 Macro jest podany wzorem:

$$F1_{macro} = \frac{1}{Q} \sum_{i=1}^{Q} \frac{2 \cdot p_i \cdot r_i}{p_i + r_i}$$

Przetwarzając wyniki jakości wnioskowania dla modeli SOTA i ChatGPT, mogliśmy obliczyć *Stratę* (*Loss*), która odzwierciedla, o ile ChatGPT jest gorszy od najlepszych dedykowanych metod. SOTA i ChatGPT oznacza tutaj wartośc miary F1 odpo-



Rysunek 12: Dane są podzielone na anotatorów (wiersze) i teksty (kolumny). Każdy z 10 zbiorów użytkowników, rozłącznych, zawiera 10% użytkowników. Podzbiory *past*, *present*, *future 1* oraz *future 2* zawierają odpowiednio 15%, 55%, 15% i 15% wszystkich tekstów, wybranych losowo. (Źródło: (Bielaniewicz i in., 2022a)).

wiednio dla modelu SOTA i GPT-3.5. *Strata* (*Loss*) jest liczona zgodnie z poniższym wzorem:

$$\text{Loss} = \frac{100\% \cdot (\text{SOTA} - \text{ChatGPT})}{\text{SOTA}}$$

Kolejna miara *Gain* kwantyfikuje jaka część całkowitej możliwej poprawy wydajności modelu referencyjnego (niespersonalizowanego NonPer) została osiągnięta przez dane rozwiązanie spersonalizowane Per, tj.:

$$\text{Gain} = \frac{100\% \cdot (\text{Per} - \text{NonPer})}{100\% - \text{NonPer}}$$

7.3.1 Badania wpływu różnych spersonalizowanych głębokich architektur predykcyjnych i modeli językowych na jakość predykcji śmieszności

W głównym scenariuszu badano wpływ różnych podejść do personalizacji, tj. efektywność wnioskowania różnych zaawansowanych głębokich architektur predykcyjnych. W szczególności dotyczy to rozwiązań zaproponowanych w niniejszej pracy: GRU (Gated Recurrent Units) (Podrozdział 6.1) oraz Wielogłowicowego mechanizmu atencji (Podrozdział 6.2). Porównano je z innymi spersonalizowanymi architekturami opracowanymi w zespole HumaNLP: HuBi-Simple, HuBi-Formula, HuBi-Medium

oraz OneHot, (Podrozdział 5). Łącznie, wszystkie ww. architektury były zwalidowane względem generalizującej Metody Referencyjnej, tj. takiej, która nie posiada żadnej wiedzy o użytkowniku a jedynie o tekście (Podrozdział 4.2). Ta ostatnia była jednak uruchamiana z zachowaniem liczności przykładów, tzn. dla każdego tekstu wnioskowanie było przeprowadzane tyle razy ilu uzytkowników go anotowało w zbiorze testowym.

Wszystkie zbiory i architektury predykcyjne były testowane na różnych modelach językowych:, tj. Random, CBOW, Skipgram, XLM-R, LaBSE, BERT, DeBERTa, MPNet. Oznacza to, że każda z uwzględnionych w badaniach architektur wykorzystuje wektorowe reprezentacje tekstów uzyskane z modelu językowego. W trakcie eksperymentów przetestowano wiele modeli językowych, aby zbadać ich wpływ na wydajność metod:

- *Random* - jest metodą losową, która opiera się na generowaniu losowych osadzeń tekstów. Została uwzględniona w celach porównawczych, by sprawdzić wpływ wiedzy z innych modeli jezykowych.

- *CBOW* (Bojanowski i in., 2017) - wariant modelu jezykowego Fasttext, którego celem jest predykcja nastepnego słowa na podstawie jego kontekstu. Model reprezentujące każde słowo jako zbiór n-gramów znakowych i uśrednia je dla każdego słowa w kontekście. Na podstawie tej uśrednionej reprezentacji przewiduje słowo centralne. Dzięki temu CBOW lepiej radzi sobie z rzadkimi słowami oraz ich odmianami, ponieważ bierze pod uwagę strukturę morfologiczną.

- *Skipgram* (Bojanowski i in., 2017) - model predykujący kontekstowe słowa na podstawie słowa centralnego. Podobnie jak w CBOW, każde słowo jest reprezentowane przez n-gramy. Skip-gram stara się przewidzieć n-gramy otaczających słów, wykorzystując n-gramy słowa centralnego. Dzięki temu, model jest bardziej efektywny w rozpoznawaniu rzadkich i nieznanych słów, co sprawia, że wariant Skip-gram modelu fasttext doskonale sprawdza się w zadaniach wymagających generowania bogatych reprezentacji dla rzadkich słów i ich odmian.

- *XLM-RoBERTa (XLM-R)* (Conneau i in., 2020) – wielojęzyczna architektura, która łączy optymalizację hiperparametrów RoBERTa z wielojęzyczną architekturą XLM, obsługującą 100 różnych języków.

- *LaBSE* (Feng i in., 2020) – model osadzeń zdań BERT, niezależny od języka, obsługujący 109 języków.

- *BERT* (Devlin i in., 2019) – model oparty na architekturze transformera, generujący dwukierunkowe reprezentacje, które uwzględniają informacje o kontekście z lewej i prawej strony we wszystkich warstwach.

- *DeBERTa* (He i in., 2020) – model rozszerzający architekturę BERT o rozszerzony dekoder maski i odseparowaną uwagę (disentangled attention).
- *MPNet* (Song i in., 2020) – model transformera, który wykorzystuje zadania wstępnego uczenia maskowanego i permutowanego modelowania języka, aby połączyć zalety maskowanego modelowania języka i permutowanego modelowania języka.

Podczas badań wykorzystano implementacje architektur transformerów oraz dedykowane tokenizatory dostarczone przez bibliotekę HuggingFace[2]. W procesie generowania reprezentacji tekstów użyto wektora zawierającego średnie wartości dla osadzeń wszystkich jego tokenów. Podejście to różni się znacząco od powszechnie stosowanej techniki wykorzystującej reprezentację specjalnego tokenu [CLS], który zawiera osadzenie dla całego tekstu. Wstępne eksperymenty wykazały, że reprezentacja tekstu oparta na średniej wartości wektorów tokenów prowadzi do znacznie lepszej wydajności niż podejście oparte wyłącznie na osadzeniu tokenu [CLS].

Proponowane modele kodują teksty za pomocą technik offline, takich jak osadzenia lub wstępnie wytrenowane modele językowe (Kocoń i in., 2021b). Taki zakres eksperymentów sprawia, że rozwiązanie jest chronione przed potencjalnymi problemami z aktualizacją danych ze względu na konieczność korzystania z metod o dużej złożoności obliczeniowej.

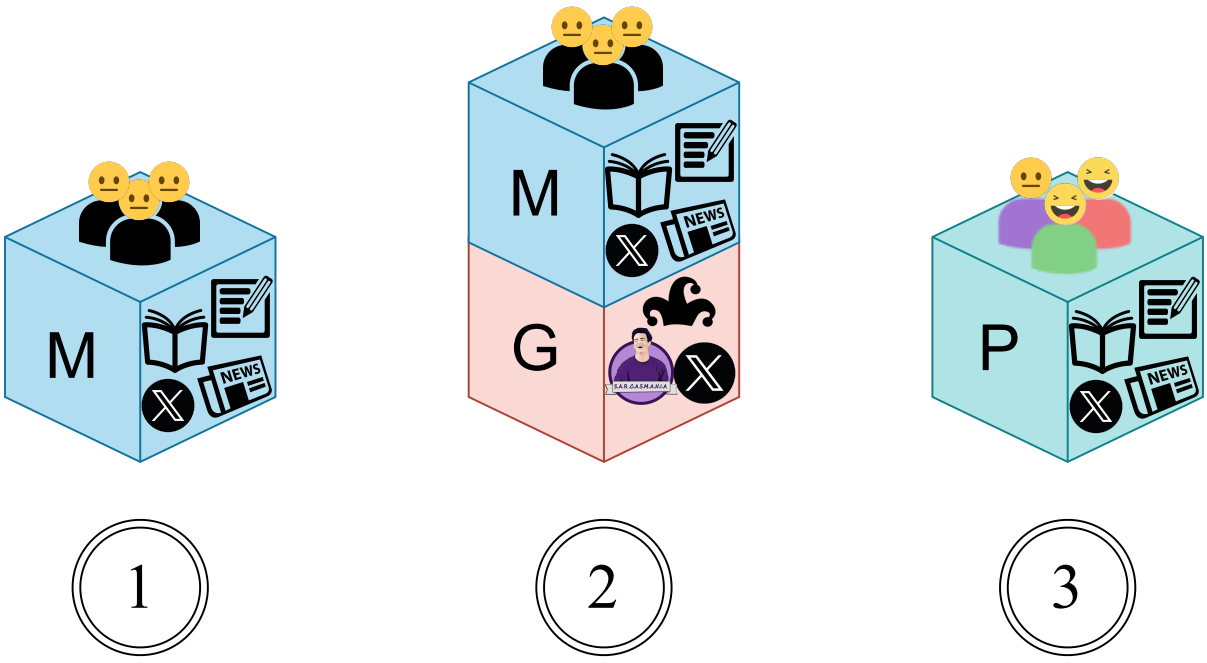
7.3.2 Badania wpływu rozmiaru zbioru uczącego na jakość predykcji śmieszności

Ten scenariusz eksperymentalny skupiał się na stopniowym zwiększaniu rozmiaru zbioru treningowego w celu zbadania tempa uczenia się charakterystyki śmieszności przez analizowane metody uczenia maszynowego. Rozpoczęto od zestawu treningowego składającego się z jednego folda w każdej iteracji walidacji krzyżowej. Następnie zwiększano rozmiar zbioru treningowego do dwóch foldów, dodając kolejny fold do tego użytego w poprzednim kroku eksperymentu. W kolejnych etapach rozmiar zbioru treningowego zwiększano o jeden fold. W ostatniej iteracji użyto ośmiu foldów jako zbioru treningowego w każdej z iteracji walidacji krzyżowej, pozostawiając dwa foldy na zbiór testowy. Scenariusz opracowano dla architektur(Model referencyjny, GRU, Attention, HuBi-Simple, HuBi-Medium, HuBi-Formula, OneHot)) i zbiorów spersonalizowanych (Cockamamie, Humicroedit, Humor, Doccano1, Doccano2, Jester), a wyniki i ich analza w kontekście tego scenariusza prezentują się w podrozdziale (Podrozdział 7.5.2 & 7.6.2).

2 https:\/\/huggingface.co\/models

7.3.3 Badania wpływu transferu uczenia na jakość predykcji śmieszności

Odczuwanie humoru na podstawie tekstu jest cechą osobowościową charakteryzującą się bardzo wysoką subiektywnością. Aby sprostać temu problemowi, proponuje się podejście fuzji danych, skoncentrowane na łączeniu anotacji użytkowników z różnych zbiorów danych dotyczących humoru. Podstawowy wariant zakłada integrację wiedzy z wielu zestawów danych humorystycznych z anotacjami użytkowników, które są agregowane poprzez głosowanie większościowe w celu zmniejszenia niepewności wynikającej z anotacji przez wielu użytkowników tekstów z różnych dziedzin i napisanych w różnych językach (pierwsza kolumna na Rys. 13). Technika ta może być również stosowana do syntezy wiedzy uzyskanej z głosowania większościowego z wiedzą ogólną na temat zabawności treści tekstowych pochodzącą z uogólnionych zbiorów danych (druga kolumna na Rys.13). Najbardziej zaawansowany wariant metody wykorzystuje oryginalne anotacje użytkowników, aby zachować indywidualne poczucie humoru każdej osoby. W ten sposób maksymalizowana jest dostępna wiedza, którą model może przyswoić w trakcie procedury treningowej.



Rysunek 13: Opracowane scenariusze fuzji danych: (1) połączone zbiory danych z głosowaniem większościowym, (2) zbiory danych z głosowaniem większościowym połączone ze zgeneralizowanymi zbiorami danych, oraz (3) połączone spersonalizowane zbiory danych. (Źródło: (Bielaniewicz i Kazienko, 2023)).

Aby ocenić metody fuzji danych zorientowanych na człowieka, wykorzystano kilka konfiguracji eksperymentalnych: (1) trenowanie modelu na próbkach z pojedynczego zbioru danych, zagregowanych za pomocą głosowania większościowego (**Majority single**), (2) łączenie próbek z wielu spersonalizowanych zbiorów danych, zagrego-

wanych za pomocą głosowania większościowego w zbiorze treningowym (pierwsza kolumna na Rys. 13, **Majority multi**), (3) łączenie próbek z wielu spersonalizowanych i zgeneralizowanych zbiorów danych podczas trenowania modelu (środkowa kolumna na Rys. 13, **Majority + Generalized multi**), (4) trenowanie modelu na indywidualnych anotacjach użytkowników z pojedynczego spersonalizowanego zbioru danych (**Personalized single**), (5) łączenie indywidualnych anotacji użytkowników ze wszystkich spersonalizowanych zbiorów danych w zbiorze treningowym (ostatnia, trzecia kolumna na Rys. 13, **Personalized multi**).

Zaprojektowano trzy główne scenariusze eksperymentalne, w których zbiory treningowe różnią się ilością zawartych w nich danych spersonalizowanych. Jeśli personalizacja była w pełni obecna, był to scenariusz **Personalized**. Jeśli jej ilość była zerowa, ta konfiguracja została nazwana **Generalization**. W przypadku opcji mieszanej, spłaszczono personalizację. W istocie nie jest możliwe łączenie zbiorów danych, gdy personalizacja jest częściowo obecna w danych, dlatego zdecydowano by przekształcić te zbiory danych przy łączeniu ich ze zgeneralizowanymi zbiorami danych. Spersonalizowane zbiory danych zostały uproszczone do postaci, która nie uwzględnia identyfikatora użytkownika, eliminując w ten sposób personalizację. Następnie, dla każdego tekstu z wieloma anotacjami, wybrano wartość, która pojawiała się najczęściej we wszystkich anotacjach, i scenariusz ten nazywa się **Majority**. W ten sposób umożliwione zostało sprawdzenie, czy dodanie ilości kosztem personalizacji mogłoby pomóc w rozumowaniu modeli.Każdy scenariusz eksperymentalny miał dwie możliwe wariacje, w zależności od liczby wykorzystanych zbiorów danych. Opcja **single** oznacza, że scenariusz eksperymentalny korzystał z pojedynczego zbioru danych, natomiast wariant **multi** obejmował kolekcję zbiorów danych odpowiednich dla każdego scenariusza. Indywidualny charakter każdego spersonalizowanego zbioru danych w kontekście anotowania zabawności stanowi wyzwanie przy próbie ich wspólnej analizy, dlatego każdy z nich został uproszczony do wartości 1, co oznacza, że coś jest zabawne, lub wartości 0, co oznacza, że tekst nie jest zabawny. Aby umożliwić wykorzystanie technik fuzji danych dla dowolnego z wybranych zbiorów danych, niezależnie od ich różnorodnych zakresów etykiet, zmapowano każdą wartość różną od zera na klasę 1 (*zabawne*). W ten sposób zachowano pierwotną informację o zabawności tekstu, która była również źródłem wiedzy o perspektywie użytkownika w przypadku spersonalizowanych zbiorów danych przedstawionych w Tab. 1.

7.4 BADANIA STATYSTYCZNE

Wszystkie przeprowadzone eksperymenty zostały powtórzone 10-krotnie w celu umożliwienia przeprowadzenia testów statystystycznych. W ramach badań przeprowadzonych na modelach uczenia maszynowego, najpierw sprawdzono założenia testu staty-

stycznego t-Studenta (Student, 1908), aby ocenić, czy możliwe jest rzetelne porównanie wyników dwóch różnych architektur sieci neuronowych uczonych na tym samym zbiorze danych.

W celu określenia, który wariant testu t-Studenta będzie odpowiedni, przeanalizowano charakter danych, aby zdecydować, czy architektury te powinny być badane testem dla prób zależnych, czy testem dla prób niezależnych. W przypadku, gdy wyniki obu architektur były powiązane (np. te same przykłady były używane w obu modelach), stosowano test t-Studenta dla prób zależnych. Natomiast, jeśli wyniki obu architektur były niezależne (np. różne zbiory uczące były używane w obu modelach), stosowano test t-Studenta dla prób niezależnych.

Aby zweryfikować założenia testu t-Studenta, zastosowano następujące metody:

- Test Shapiro-Wilka (Shapiro i Wilk, 1965) – użyty do sprawdzenia, czy rozkłady wyników dla obu architektur są normalne.
- Test na równość wariancji (Bartlett, 1937) – aby upewnić się, że wariancje wyników dla obu grup są homogeniczne.
- Sprawdzenie niezależności obserwacji – aby potwierdzić, że poszczególne wyniki w ramach każdej z próbek są od siebie niezależne.

Jeżeli wszystkie założenia testu t-Studenta zostały spełnione, przeprowadzono odpowiedni test t-Studenta, zależny lub niezależny, w zależności od charakteru danych. W przypadku, gdy założenia testu t-Studenta nie były spełnione, zamiast tego zastosowano testy nieparametryczne, takie jak test Wilcoxona (dla prób zależnych) (Wilcoxon, 1945) lub test Manna-Whitneya (dla prób niezależnych) (Mann i Whitney, 1947), które nie wymagają spełnienia założeń dotyczących normalności rozkładu oraz równości wariancji.

W przeprowadzonych badaniach uczenia maszynowego, które obejmowały liczne testy statystyczne, zastosowano poprawkę Bonferroniego (Dunn, 1961) w celu kontrolowania błędu typu I. W kontekście porównań wielu par grup, ryzyko fałszywych pozytywów, czyli błędnego uznania, że różnice między grupami są istotne, gdy w rzeczywistości są przypadkowe, zostało znacząco zwiększone. Aby zminimalizować to ryzyko, poziom istotności (α) został skorygowany na podstawie liczby przeprowadzanych testów. Dzięki zastosowaniu poprawki Bonferroniego, zapewniono bardziej rygorystyczne podejście do analizy wyników, co zwiększyło wiarygodność wniosków wyciągniętych z badań.

7.5 WYNIKI

Rozdział ten prezentuje szczegółowe rezultaty przeprowadzonych badań nad wpływem różnych modeli językowych na jakość predykcji śmieszności, wpływem rozmiaru zbioru uczącego na jakość predykcji śmieszności oraz wpływu transferu uczenia na jakość predykcji śmieszności. Zawarte tu analizy obejmują ocenę wydajności modeli w różnych zadaniach, takich jak klasyfikacja czy regresja. Wyniki eksperymentów porównano na podstawie miar jakościowych, takich jak Macro F1 i Mean Squared Error (MSE), dla różnych zestawów danych humorystycznych. Każdy podrozdział szczegółowo omawia uzyskane wyniki i przedstawia wnioski dotyczące skuteczności testowanych modeli.

7.5.1 Badania wpływu różnych spersonalizowanych głębokich architektur predykcyjnych i modeli językowych na jakość predykcji śmieszności

W badaniach nad wpływem różnych modeli językowych na jakość predykcji śmieszności celem eksperymentów było porównanie wyników uzyskanych przez różne modele, w tym szczególnie zaawansowane głębokie architektury predykcyjne, takie jak zaproponowane w niniejszej pracy GRU (Gated Recurrent Units) oraz Wielogłowicowy mechanizm atencji, jak również spersonalizowane architektury opracowane w zespole HumaNLP (HuBi-Simple, HuBi-Formula, HuBi-Medium oraz OneHot). Wszystkie architektury były porównane z generalizującą Metodą Referencyjną bez wiedzy o użytkowniku.

Wszystkie zbiory były testowane na różnych modelach językowych:, tj. Random, CBOW, Skipgram, XLM-R, LaBSE, BERT, DeBERTa, MPNet.

Wyniki miar Macro F1, przedstawione w Tab. 13, pokazują, że dla zbioru Cockamamie modele GRU i Attention osiągnęły znacząco wyższe rezultaty niż model bazowy. Wartość Macro F1 wzrosła istotnie w porównaniu do modelu referencyjnego, co wskazuje na lepsze zdolności tych modeli do uchwycenia niuansów humorystycznych. Podobnie jak w przypadku zbioru Cockamamie, eksperymenty na zbiorze Humicroedit wykazały przewagę zaawansowanych modeli, co odzwierciedlają wartości Macro F1 przedstawione w Tab. 14. Modele GRU i Attention wyraźnie poprawiły wyniki w porównaniu z metodami referencyjnymi, co sugeruje, że ich zdolność do modelowania kontekstu i struktury językowej humoru jest kluczowa w tego typu zadaniach. Zbiór danych Humor również potwierdził efektywność zaawansowanych architektur. Wyniki przedstawione w Tab. 15 pokazują, że modele GRU i Attention ponownie uzyskały lepsze rezultaty niż tradycyjne podejścia, co potwierdza ich uniwersalność i zdolność do poprawy jakości predykcji w różnych zbiorach danych humorystycznych.

Zbiory danych Doccano1 oraz Doccano2, których wyniki przedstawiono odpowiednio w Tab. 16 oraz Tab. 17, również wykazały pozytywny wpływ bardziej złożonych modeli na wyniki predykcji. Chociaż różnice w wynikach miary Macro F1 były mniej wyraźne niż w poprzednich zbiorach, nadal można zaobserwować, że modele GRU i Attention przewyższały klasyczne podejścia. Zbiór Jester został wykorzystany w zadaniu regresji, gdzie oceniano jakość predykcji za pomocą miary Mean Squared Error (MSE). Wyniki przedstawione w Tab. 18 wskazują, że modele GRU i Attention poprawiły wyniki, jednak korzyści te były mniej wyraźne niż w przypadku miar klasyfikacyjnych, co sugeruje, że zaawansowane modele mogą nie przynosić tak znaczących korzyści w zadaniach regresyjnych jak w klasyfikacyjnych. Badania potwierdziły, że zaawansowane architektury, takie jak GRU oraz wielogłowicowy mechanizm atencji, mają istotny wpływ na poprawę jakości predykcji śmieszności w zadaniach klasyfikacyjnych. Największe korzyści zaobserwowano w zbiorach danych Cockamamie i Humicroedit, gdzie różnice w miarach Macro F1 były najbardziej widoczne. Mimo że w zadaniach regresyjnych, takich jak w przypadku zbioru Jester, wyniki były mniej spektakularne, zastosowanie tych modeli nadal przyniosło pozytywne efekty.

7.5.2 Badania wpływu rozmiaru zbioru uczącego na jakość predykcji śmieszności

W ramach badań nad wpływem rozmiaru zbioru uczącego na jakość predykcji śmieszności przeprowadzono szereg eksperymentów. Ich analiza obejmowała ocenę wyników dla różnych liczby foldów uczących, co pozwoliło na zbadanie, jak zmiana rozmiaru zbioru danych wpływa na zdolność modeli do skutecznego przewidywania śmieszności. Dla zbioru Cockamamie, wyniki miary Macro F1 w zależności od liczby foldów uczących przedstawione w Tab. 5 pokazują, że wraz ze zwiększaniem liczby foldów uczących jakość predykcji systematycznie rośnie. Im większy rozmiar zbioru uczącego, tym wyższa była wartość Macro F1, co wskazuje na lepsze przystosowanie modeli do rozpoznawania śmieszności.

Wyniki dla miary precyzji, prezentowane w Tab. 6, pokazują podobną tendencję – większe zbiory danych prowadziły do poprawy precyzji modeli. Z kolei wartości miary kompletności (recall) z Tab. 7 również wzrastały wraz z większą liczbą foldów uczących, co świadczy o lepszej zdolności modeli do wykrywania wszystkich przypadków śmieszności.

Zbiór danych Humicroedit również wykazał pozytywny wpływ zwiększenia rozmiaru zbioru uczącego na jakość predykcji. Wyniki Macro F1, przedstawione w Tab. 8, wskazują na znaczną poprawę wydajności modeli wraz ze wzrostem liczby foldów. Modele uczące się na większych zestawach danych były w stanie lepiej rozpoznawać wzorce humorystyczne, co przełożyło się na lepsze wyniki w klasyfikacji śmieszności.

Dla zbioru Humor wyniki miary Macro F1, które znajdują się w Tab. 9, pokazują, że większy rozmiar zbioru uczącego miał pozytywny wpływ na jakość predykcji. Modele trenowane na większych zbiorach danych uzyskiwały lepsze rezultaty, co potwierdza, że dostęp do większej ilości danych umożliwia lepsze generalizowanie i rozpoznawanie niuansów związanych z humorem.

Zbiory Doccano1 oraz Doccano2, których wyniki przedstawiono w Tab. 10 oraz Tab. 11, także potwierdziły znaczący wpływ rozmiaru zbioru uczącego na jakość predykcji. W obu przypadkach większe zbiory danych pozwoliły modelom na uzyskanie lepszych wyników miary Macro F1, co świadczy o lepszej zdolności do rozpoznawania śmieszności w przypadku większej liczby danych treningowych.

Dla zbioru Jester, który został poddany ocenie za pomocą miary Mean Squared Error (MSE), wyniki przedstawione w Tab. 12 pokazują, że w zadaniach regresyjnych zwiększenie liczby foldów uczących również prowadziło do obniżenia błędu predykcji. Modele trenowane na większych zbiorach danych były bardziej precyzyjne, co przełożyło się na mniejsze wartości MSE.

Badania wykazały, że rozmiar zbioru uczącego ma istotny wpływ na jakość predykcji śmieszności we wszystkich analizowanych zbiorach danych. Wraz ze wzrostem liczby foldów uczących poprawiała się zarówno miara Macro F1, jak i precyzja oraz kompletność, co świadczy o tym, że większe zbiory danych umożliwiają modelom lepsze rozpoznawanie wzorców humorystycznych. Również w zadaniach regresyjnych, jak w przypadku zbioru Jester, większe zbiory danych prowadziły do poprawy wyników, co potwierdza uniwersalny wpływ rozmiaru zbioru na jakość predykcji.

7.5.3 Badania wpływu transferu uczenia na jakość predykcji śmieszności

W niniejszych badaniach transfer uczenia został zastosowany kontekście predykcji śmieszności, aby zbadać, czy wykorzystanie wiedzy zdobytej na jednym zestawie danych może poprawić wyniki na innych zbiorach danych.

Wyniki dla zbioru Cockamamie przedstawione w Tab. 19 wykazują, że zastosowanie transferu uczenia znacząco poprawiło jakość predykcji śmieszności. Przeniesienie wiedzy z większych zbiorów danych pozwoliło modelom uzyskać wyższe wartości miary Macro F1 niż w przypadku trenowania modeli od podstaw. Transfer uczenia zwiększył zdolność modelu do rozpoznawania bardziej subtelnych aspektów humoru, co przełożyło się na wyraźną poprawę wyników. Podobne efekty zaobserwowano dla zbioru Humicroedit, którego wyniki zamieszczono w Tab. 20. Transfer uczenia przyczynił się do znaczącej poprawy wartości Macro F1, szczególnie w przypadku mniejszych zbiorów danych, które same w sobie mogłyby nie dostarczyć wystarczającej ilości informacji do skutecznego trenowania modeli. Wykorzystanie wiedzy z innych domen umożliwiło lepszą generalizację modeli, co przełożyło się na

Tabela 5: Wartości miary macro F-1 na przestrzeni wszystkich foldów podczas uczenia modeli na zbiorze danych Cockamamie.

Architektura \ Foldy uczące	1	1-2	1-3	1-4	1-5	1-6	1-7	1-8
Metoda Referencyjna	49.88 ± 0.85	50.18 ± 0.68	50.38 ± 0.56	50.50 ± 0.59	50.50 ± 0.49	50.56 ± 0.74	50.82 ± 0.39	50.94 ± 0.97
GRU	**51.57** ± 0.75	**52.24** ± 1.28	**52.64** ± 0.59	**52.87** ± 1.27	53.47 ± 1.17	53.48 ± 1.29	53.67 ± 0.93	53.78 ± 0.96
Attention	51.32 ± 0.77	52.01 ± 0.68	52.52 ± 0.85	52.72 ± 1.16	**53.50** ± 1.45	**53.78** ± 1.26	**53.87** ± 1.49	**54.12** ± 1.07
HuBi-Simple	50.53 ± 1.02	50.72 ± 0.95	50.90 ± 1.09	50.92 ± 0.97	50.97 ± 1.18	51.08 ± 0.97	51.18 ± 1.00	51.33 ± 1.03
HuBi-Formula	50.79 ± 1.05	50.83 ± 1.22	51.08 ± 1.64	51.10 ± 1.70	51.57 ± 1.34	52.33 ± 1.23	52.65 ± 1.78	52.71 ± 1.53
HuBi-Medium	49.98 ± 1.44	49.99 ± 0.88	49.99 ± 1.30	50.02 ± 1.09	50.03 ± 0.94	50.09 ± 0.59	50.10 ± 0.82	51.46 ± 0.86
OneHot	51.12 ± 0.37	51.48 ± 0.48	51.74 ± 1.25	52.01 ± 1.17	52.07 ± 0.35	52.58 ± 1.31	52.62 ± 0.34	53.04 ± 1.74

dokładniejsze przewidywanie śmieszności. W zbiorze Humor transfer uczenia również wykazał swoją wartość, co ilustrują wyniki w Tab. 21. Przeniesienie wiedzy z innych humorystycznych zbiorów danych pozwoliło na poprawę jakości predykcji śmieszności, zwłaszcza w tekstach o bardziej złożonych strukturach humorystycznych. Zbiór Doccano1 również wykazał wyraźną poprawę wyników dzięki zastosowaniu transferu uczenia, co ilustruje Tab. 22. Modele z transferem uczenia osiągnęły wyższe wartości Macro F1 w porównaniu do modeli trenowanych na samym zbiorze Doccano1. Wzrost precyzji i dokładności predykcji wynikał z przeniesienia wiedzy z innych zbiorów, co pozwoliło modelom lepiej radzić sobie z różnorodnymi formami humoru obecnymi w zbiorze Doccano1. Dla zbioru Doccano2 (Tab. 23), transfer uczenia również przyniósł pozytywne efekty. Modele trenowane z wykorzystaniem wiedzy z innych zestawów danych uzyskały wyższe wartości Macro F1, co wskazuje na lepsze rozpoznawanie humoru w tekstach o bardziej specyficznej strukturze. Przeniesienie wiedzy z większych zbiorów pomogło modelom radzić sobie z bardziej skomplikowanymi i specyficznymi kontekstami humorystycznymi, co przełożyło się na poprawę jakości predykcji. Zbiór Jester stanowił wyjątkowy przypadek, ponieważ zastosowano tutaj zadanie regresji, a nie klasyfikacji. W Tab. 24 zaprezentowano wyniki miary Mean Squared Error (MSE), które pokazują, że transfer uczenia również w zadaniach regresyjnych poprawił dokładność modeli. Choć różnice nie były tak duże jak w przypadku zadań klasyfikacyjnych, transfer uczenia nadal przyniósł korzyści, obniżając wartość MSE i poprawiając ogólną jakość predykcji. Wyniki badań jednoznacznie potwierdzają, że transfer uczenia znacząco wpływa na poprawę jakości predykcji śmieszności. Modele, które korzystały z wiedzy przeniesionej z innych zbiorów danych, uzyskiwały lepsze wyniki w miarach Macro F1 oraz MSE w porównaniu do modeli trenowanych od podstaw. Największe korzyści zaobserwowano w przypadkach, gdy modele trenowane były na mniejszych zbiorach danych, gdzie bez transferu uczenia ich zdolności predykcyjne byłyby ograniczone.

Tabela 6: Wartości miary precyzji na przestrzeni wszystkich foldów podczas uczenia modeli na zbiorze danych Cockamamie.

Architektura \ Foldy uczące	1	1-2	1-3	1-4	1-5	1-6	1-7	1-8
Metoda Referencyjna	49.54 ± 0.73	50.03 ± 0.63	50.51 ± 0.72	51.03 ± 0.68	51.53 ± 0.69	52.03 ± 0.57	52.56 ± 0.53	53.01 ± 0.48
GRU	**51.56** ± 0.83	**52.49** ± 0.76	**53.47** ± 0.86	**54.45** ± 0.97	**55.49** ± 1.04	**56.53** ± 1.08	**57.51** ± 1.13	**58.54** ± 1.19
Attention	51.07 ± 0.82	52.04 ± 0.96	52.50 ± 0.93	53.07 ± 1.02	54.01 ± 1.07	55.02 ± 1.10	56.01 ± 1.18	57.03 ± 1.22
HuBi-Simple	49.53 ± 0.75	50.01 ± 0.79	50.56 ± 0.84	51.02 ± 0.83	51.50 ± 0.94	52.02 ± 0.99	52.56 ± 1.08	53.02 ± 1.03
HuBi-Formula	50.03 ± 0.77	50.52 ± 0.83	51.04 ± 0.89	51.53 ± 0.92	52.03 ± 0.94	52.52 ± 1.00	53.02 ± 1.03	53.51 ± 1.13
HuBi-Medium	49.04 ± 0.72	49.51 ± 0.73	50.04 ± 0.83	50.58 ± 0.88	51.01 ± 0.93	51.59 ± 0.98	52.03 ± 1.04	52.56 ± 1.03
OneHot	51.01 ± 0.65	51.53 ± 0.72	52.04 ± 0.79	52.7 ± 0.82	53.05 ± 0.86	53.52 ± 0.94	54.03 ± 0.95	54.51 ± 1.03

Tabela 7: Wartości miary kompletności na przestrzeni wszystkich foldów podczas uczenia modeli na zbiorze danych Cockamamie.

Architektura \ Foldy uczące	1	1-2	1-3	1-4	1-5	1-6	1-7	1-8
Metoda Referencyjna	50.03 ± 0.72	50.51 ± 0.70	51.03 ± 0.72	51.54 ± 0.65	52.03 ± 0.62	52.51 ± 0.54	53.03 ± 0.51	53.53 ± 0.42
GRU	50.57 ± 0.88	51.51 ± 0.86	52.53 ± 0.94	53.51 ± 0.97	**54.58** ± 1.02	**55.53** ± 1.07	**56.58** ± 1.12	57.55 ± 1.19
Attention	**51.56** ± 0.84	**52.53** ± 0.98	**53.08** ± 0.94	**53.53** ± 1.07	54.51 ± 1.08	55.50 ± 1.19	56.53 ± 1.14	**57.56** ± 1.24
HuBi-Simple	50.09 ± 0.74	50.57 ± 0.76	51.07 ± 0.83	51.52 ± 0.83	52.04 ± 0.92	52.50 ± 0.94	53.08 ± 1.05	53.59 ± 1.07
HuBi-Formula	50.54 ± 0.73	51.07 ± 0.86	51.52 ± 0.84	52.03 ± 0.97	52.52 ± 0.96	53.01 ± 1.02	53.54 ± 1.06	54.08 ± 1.10
HuBi-Medium	49.54 ± 0.73	50.05 ± 0.77	50.51 ± 0.82	51.06 ± 0.88	51.50 ± 0.96	52.03 ± 0.95	52.4 ± 1.54	53.03 ± 1.09
OneHot	51.54 ± 0.63	52.07 ± 0.79	52.52 ± 0.70	53.04 ± 0.85	53.59 ± 0.83	54.04 ± 0.92	54.52 ± 0.95	55.03 ± 1.04

Tabela 8: Wartości miary macro F-1 na przestrzeni wszystkich foldów podczas uczenia modeli na zbiorze danych Humicroedit.

Architektura \ Foldy uczące	1	1-2	1-3	1-4	1-5	1-6	1-7	1-8
Metoda Referencyjna	47.32 ± 2.51	47.54 ± 1.76	47.67 ± 2.44	47.73 ± 2.67	47.74 ± 2.60	48.00 ± 2.58	48.04 ± 2.12	48.26 ± 2.55
GRU	49.34 ± 0.90	**49.75** ± 1.23	50.02 ± 0.80	**50.34** ± 1.08	50.51 ± 0.94	50.60 ± 1.19	**50.97** ± 1.24	**51.14** ± 0.56
Attention	**49.42** ± 1.16	49.72 ± 0.62	**50.20** ± 0.48	50.29 ± 0.82	**50.57** ± 0.98	**50.73** ± 1.03	50.77 ± 0.24	50.97 ± 0.69
HuBi-Simple	48.44 ± 1.57	48.56 ± 0.95	48.95 ± 1.24	49.10 ± 0.49	49.48 ± 1.47	49.68 ± 1.16	49.73 ± 0.69	49.98 ± 1.12
HuBi-Formula	49.12 ± 0.89	49.13 ± 0.29	49.39 ± 1.20	49.55 ± 1.14	49.94 ± 1.00	50.19 ± 0.94	50.67 ± 0.89	50.72 ± 0.52
HuBi-Medium	40.95 ± 0.94	41.11 ± 0.48	41.26 ± 1.02	41.59 ± 0.28	41.77 ± 0.30	42.08 ± 0.78	42.13 ± 0.36	49.38 ± 0.96
OneHot	49.17 ± 0.50	49.26 ± 0.47	49.77 ± 0.55	49.95 ± 0.82	50.03 ± 0.85	50.05 ± 0.50	50.57 ± 0.78	50.58 ± 0.71

Tabela 9: Wartości miary macro F-1 na przestrzeni wszystkich foldów podczas uczenia modeli na zbiorze danych Humor.

Architektura \ Foldy uczące	1	1-2	1-3	1-4	1-5	1-6	1-7	1-8
Metoda Referencyjna	68.92 ± 1.23	69.24 ± 0.57	70.30 ± 0.53	70.61 ± 1.73	71.24 ± 1.28	71.34 ± 1.23	72.23 ± 0.55	72.97 ± 1.33
GRU	**71.82** ± 0.73	72.10 ± 0.58	72.13 ± 0.87	72.26 ± 0.49	**72.94** ± 0.72	72.97 ± 0.96	**73.36** ± 0.49	74.03 ± 0.93
Attention	71.75 ± 0.37	**72.17** ± 0.69	**72.68** ± 0.62	**72.73** ± 0.25	72.90 ± 0.77	72.98 ± 0.56	73.23 ± 0.75	**74.12** ± 0.25
HuBi-Simple	71.37 ± 1.50	71.81 ± 0.91	71.89 ± 1.19	72.39 ± 1.08	72.74 ± 1.21	**73.05** ± 1.27	73.25 ± 0.67	73.73 ± 0.92
HuBi-Formula	71.53 ± 1.01	71.55 ± 1.50	71.74 ± 1.65	71.98 ± 0.43	71.99 ± 0.92	72.62 ± 0.55	72.97 ± 1.19	73.37 ± 0.59
HuBi-Medium	71.37 ± 0.85	71.61 ± 1.03	71.72 ± 1.08	72.17 ± 0.89	72.22 ± 0.89	72.66 ± 0.90	72.73 ± 1.08	73.08 ± 0.83
OneHot	71.12 ± 1.37	71.19 ± 1.00	71.76 ± 1.08	71.98 ± 1.78	72.08 ± 1.09	72.30 ± 0.54	72.66 ± 1.33	72.99 ± 0.75

Tabela 10: Wartości miary macro F-1 na przestrzeni wszystkich foldów podczas uczenia modeli na zbiorze danych Doccano 1.

Architektura \ Foldy uczące	1	1-2	1-3	1-4	1-5	1-6	1-7	1-8
Metoda Referencyjna	44.68 ± 2.64	45.19 ± 2.54	45.64 ± 1.89	45.87 ± 1.82	46.61 ± 1.45	46.70 ± 2.52	47.68 ± 2.03	47.96 ± 2.30
GRU	46.97 ± 1.81	**48.05** ± 1.23	**48.40** ± 1.77	48.68 ± 1.84	48.78 ± 1.05	49.23 ± 1.60	49.56 ± 1.03	50.32 ± 1.25
Attention	**47.02** ± 0.48	47.81 ± 1.28	48.14 ± 1.58	**48.92** ± 1.29	49.25 ± 0.78	**49.63** ± 0.71	**50.52** ± 1.16	**50.69** ± 0.65
HuBi-Simple	46.79 ± 1.80	47.74 ± 2.65	48.20 ± 1.46	48.63 ± 2.09	**49.30** ± 1.87	49.61 ± 1.30	49.90 ± 1.51	49.99 ± 1.44
HuBi-Formula	43.87 ± 0.90	44.15 ± 1.99	44.39 ± 1.21	45.05 ± 1.79	47.98 ± 2.69	48.05 ± 2.04	48.51 ± 2.97	49.34 ± 2.66
HuBi-Medium	43.87 ± 1.43	44.46 ± 1.90	44.90 ± 0.97	45.00 ± 2.03	45.50 ± 0.80	45.56 ± 1.17	46.12 ± 0.89	48.71 ± 0.68
OneHot	46.32 ± 2.13	46.66 ± 2.00	46.83 ± 1.10	47.52 ± 1.82	48.09 ± 1.92	48.86 ± 1.48	49.42 ± 2.28	49.82 ± 2.32

Tabela 11: Wartości miary macro F-1 na przestrzeni wszystkich foldów podczas uczenia modeli na zbiorze danych Doccano 2.

Architektura \ Foldy uczące	1	1-2	1-3	1-4	1-5	1-6	1-7	1-8
Metoda Referencyjna	48.15 ± 1.22	48.41 ± 0.99	48.64 ± 0.89	48.92 ± 1.18	49.39 ± 1.36	49.61 ± 1.40	49.98 ± 1.16	50.10 ± 1.39
GRU	61.42 ± 1.62	**62.13** ± 0.82	62.20 ± 1.38	62.58 ± 0.97	**63.17** ± 1.37	**63.25** ± 1.37	63.60 ± 0.86	63.85 ± 0.81
Attention	**61.65** ± 1.25	61.99 ± 1.12	**62.47** ± 1.34	**62.74** ± 1.00	63.04 ± 0.90	63.24 ± 1.00	**63.74** ± 1.14	**64.13** ± 0.87
HuBi-Simple	58.16 ± 1.83	58.25 ± 1.86	58.89 ± 2.15	59.52 ± 1.91	60.21 ± 2.43	60.96 ± 2.40	61.37 ± 1.56	62.40 ± 2.20
HuBi-Formula	57.86 ± 0.86	58.32 ± 1.24	58.60 ± 1.14	58.67 ± 1.34	59.47 ± 1.17	59.82 ± 1.24	60.25 ± 1.14	60.89 ± 0.96
HuBi-Medium	50.78 ± 0.86	51.27 ± 0.69	51.35 ± 0.56	51.42 ± 0.79	51.86 ± 0.61	52.10 ± 0.95	52.44 ± 0.72	52.54 ± 0.71
OneHot	61.27 ± 1.64	61.70 ± 1.14	61.85 ± 2.01	61.98 ± 2.06	62.07 ± 1.00	62.37 ± 1.59	62.72 ± 1.90	63.06 ± 1.52

Tabela 12: Wartości miary mean squared error (MSE) na przestrzeni wszystkich foldów podczas uczenia modeli na zbiorze danych Jester.

Architektura \ Foldy uczące	1	1-2	1-3	1-4	1-5	1-6	1-7	1-8
Metoda Referencyjna	0.151 ± 0.012	0.143 ± 0.011	0.132 ± 0.010	0.121 ± 0.030	0.101 ± 0.020	0.092 ± 0.031	0.083 ± 0.021	0.071 ± 0.013
GRU	0.071 ± 0.020	0.073 ± 0.020	0.074 ± 0.021	0.074 ± 0.021	**0.060** ± 0.012	0.062 ± 0.021	0.062 ± 0.011	0.060 ± 0.021
Attention	**0.061** ± 0.030	**0.062** ± 0.021	0.063 ± 0.031	**0.062** ± 0.023	**0.060** ± 0.010	**0.051** ± 0.022	**0.053** ± 0.020	**0.051** ± 0.014
HuBi-Simple	0.081 ± 0.002	0.080 ± 0.011	0.080 ± 0.002	0.073 ± 0.001	0.070 ± 0.001	0.071 ± 0.013	0.061 ± 0.003	0.062 ± 0.011
HuBi-Formula	0.131 ± 0.023	0.111 ± 0.012	0.113 ± 0.001	0.103 ± 0.031	0.082 ± 0.011	0.083 ± 0.030	0.071 ± 0.010	0.061 ± 0.031
HuBi-Medium	0.091 ± 0.011	0.092 ± 0.001	0.092 ± 0.001	0.091 ± 0.002	0.080 ± 0.013	0.070 ± 0.011	0.072 ± 0.001	0.061 ± 0.001
OneHot	0.071 ± 0.020	0.071 ± 0.013	**0.062** ± 0.013	0.064 ± 0.021	0.062 ± 0.023	0.061 ± 0.010	0.061 ± 0.012	0.064 ± 0.010

Tabela 13: Wartość miary macro F1 oraz dla klasy reprezentującej kategorię śmieszność w zależności od modelu wykorzystanego do wygenerowania reprezentacji wektorowych tekstów dla zbioru danych Cockamamie.

	Miara	Random	CBOW	Skipgram	BERT	DeBERTa	XLM-R	LaBSE	MPNet
Metoda Referencyjna	F1	41.03 ± 03.97	40.32 ± 02.39	40.52 ± 01.86	41.19 ± 01.22	43.68 ± 03.10	39.26 ± 01.24	50.94 ± 0.97	41.52 ± 02.72
	F1-Funny	09.57 ± 08.17	05.55 ± 03.32	04.52 ± 02.81	04.83 ± 03.81	10.38 ± 06.62	02.53 ± 00.95	10.79 ± 05.12	04.62 ± 03.90
GRU	F1	47.23 ± 0.99	47.92 ± 0.93	49.04 ± 1.29	49.74 ± 0.98	51.25 ± 1.31	51.82 ± 1.03	53.78 ± 0.96	53.50 ± 0.95
	F1-Funny	**19.84** ± 3.03	21.82 ± 1.34	24.03 ± 4.18	25.86 ± 4.14	28.38 ± 4.56	29.49 ± 2.46	**35.61** ± 2.93	31.17 ± 3.46
Attention	F1	**47.69** ± 1.50	48.31 ± 1.75	49.41 ± 0.65	**50.34** ± 1.17	51.06 ± 1.57	**52.36** ± 1.80	**54.12** ± 1.07	**54.01** ± 1.82
	F1-Funny	17.84 ± 3.83	21.99 ± 4.19	24.94 ± 2.65	**29.30** ± 4.28	**31.81** ± 3.52	**32.56** ± 1.96	35.12 ± 1.84	**36.17** ± 2.24
HuBi-Simple	F1	46.50 ± 05.89	49.37 ± 05.67	49.43 ± 04.99	47.72 ± 06.07	51.33 ± 1.03	45.81 ± 05.05	50.99 ± 04.83	50.19 ± 05.84
	F1-Funny	16.59 ± 14.38	22.67 ± 13.04	21.67 ± 09.36	19.71 ± 13.82	26.25 ± 11.33	14.87 ± 10.97	16.68 ± 11.35	23.22 ± 11.82
HuBi-Formula	F1	46.21 ± 05.87	47.03 ± 04.88	47.02 ± 03.67	49.08 ± 02.50	49.00 ± 05.05	43.65 ± 03.52	52.71 ± 1.53	49.12 ± 05.28
	F1-Funny	17.09 ± 13.61	17.07 ± 09.71	16.83 ± 07.15	21.44 ± 04.96	22.10 ± 09.64	10.43 ± 06.76	23.62 ± 05.68	21.37 ± 10.87
HuBi-Medium	F1	42.21 ± 01.14	**50.13** ± 0.86	**50.02** ± 03.94	43.33 ± 01.38	50.00 ± 02.33	43.98 ± 04.76	51.46 ± 03.74	48.42 ± 06.37
	F1-Funny	10.01 ± 02.74	**28.02** ± 08.93	**25.06** ± 08.40	17.53 ± 01.68	12.60 ± 05.06	10.36 ± 04.63	15.53 ± 08.01	19.76 ± 09.79
OneHot	F1	46.66 ± 03.81	48.34 ± 03.81	48.97 ± 04.14	49.62 ± 02.88	**53.04** ± 1.74	47.00 ± 04.31	51.24 ± 02.96	50.21 ± 04.51
	F1-Funny	17.88 ± 10.08	19.77 ± 08.91	20.82 ± 07.78	23.36 ± 05.93	26.28 ± 08.07	18.11 ± 08.81	22.20 ± 07.95	23.56 ± 09.18

Tabela 14: Wartość miary macro F1 oraz dla klasy reprezentującej kategorię śmieszność w zależności od modelu wykorzystanego do wygenerowania reprezentacji wektorowych tekstów dla zbioru danych Humicroedit.

	Miara	Random	CBOW	Skipgram	BERT	DeBERTa	XLM-R	LaBSE	MPNet
Metoda Referencyjna	F1	40.23 ± 2.89	47.50 ± 4.66	36.84 ± 3.40	40.54 ± 3.96	37.75 ± 2.10	42.17 ± 1.54	48.26 ± 2.55	42.60 ± 4.80
	F1-Funny	22.46 ± 4.19	19.35 ± 2.15	17.89 ± 3.73	20.84 ± 3.41	23.97 ± 4.32	18.56 ± 2.85	25.05 ± 2.37	24.92 ± 3.61
GRU	F1	46.77 ± 3.13	47.93 ± 1.70	**49.17** ± 1.88	**49.74** ± 2.37	44.84 ± 2.06	48.35 ± 1.94	**51.14** ± 0.56	**47.51** ± 1.85
	F1-Funny	29.06 ± 1.79	29.84 ± 3.16	**33.30** ± 1.38	25.05 ± 3.18	26.79 ± 2.60	27.33 ± 2.96	**35.89** ± 3.27	**32.25** ± 3.24
Attention	F1	43.51 ± 3.05	48.35 ± 2.25	46.76 ± 3.13	44.33 ± 1.69	49.30 ± 3.90	42.84 ± 1.99	50.97 ± 0.69	46.28 ± 2.72
	F1-Funny	26.05 ± 2.38	**35.01** ± 2.32	27.98 ± 1.86	**29.39** ± 1.95	**31.87** ± 1.13	30.81 ± 1.24	33.04 ± 3.24	28.32 ± 1.89
HuBi-Simple	F1	46.09 ± 3.41	**49.02** ± 2.96	44.09 ± 1.90	43.11 ± 2.09	43.45 ± 3.02	42.52 ± 3.21	49.98 ± 1.12	44.39 ± 3.81
	F1-Funny	27.10 ± 2.96	24.05 ± 2.27	30.39 ± 1.80	28.01 ± 2.52	26.79 ± 2.72	31.37 ± 2.19	32.17 ± 3.18	26.28 ± 2.95
HuBi-Formula	F1	**49.31** ± 3.60	47.59 ± 1.52	44.41 ± 1.87	47.18 ± 2.17	47.42 ± 3.05	47.64 ± 1.59	50.72 ± 0.52	44.45 ± 3.87
	F1-Funny	**30.39** ± 4.09	31.99 ± 3.06	27.07 ± 4.02	28.85 ± 3.95	26.01 ± 4.43	30.62 ± 2.39	32.05 ± 4.46	27.63 ± 3.02
HuBi-Medium	F1	40.64 ± 2.89	47.87 ± 2.55	40.18 ± 2.06	42.01 ± 3.01	39.02 ± 2.41	42.74 ± 2.08	49.38 ± 0.96	43.25 ± 2.65
	F1-Funny	23.05 ± 4.82	22.16 ± 1.66	23.80 ± 4.29	23.23 ± 2.38	23.99 ± 1.14	25.01 ± 2.04	26.57 ± 1.72	25.06 ± 1.14
OneHot	F1	44.32 ± 2.55	47.71 ± 2.89	47.13 ± 3.70	46.00 ± 4.05	**49.89** ± 1.07	**48.47** ± 2.52	50.58 ± 0.71	44.17 ± 1.98
	F1-Funny	30.88 ± 4.40	24.05 ± 4.33	26.31 ± 4.07	27.00 ± 1.63	30.39 ± 2.24	30.90 ± 1.49	32.01 ± 4.59	28.38 ± 2.29

Tabela 15: Wartość miary macro F1 oraz dla klasy reprezentującej kategorię śmieszność w zależności od modelu wykorzystanego do wygenerowania reprezentacji wektorowych tekstów dla zbioru danych Humor.

	Miara	Random	CBOW	Skipgram	BERT	DeBERTa	XLM-R	LaBSE	MPNet
Metoda Referencyjna	F1	66.70 ± 3.08	64.05 ± 4.08	67.88 ± 4.67	64.41 ± 1.84	64.25 ± 3.63	66.50 ± 2.71	72.97 ± 1.33	62.23 ± 1.64
	F1-Funny	19.09 ± 4.45	19.48 ± 3.45	18.99 ± 2.16	24.07 ± 2.91	19.96 ± 2.30	19.54 ± 4.44	26.21 ± 4.87	22.07 ± 3.31
GRU	F1	67.05 ± 2.87	**71.70** ± 1.56	71.24 ± 3.43	67.88 ± 1.72	69.20 ± 2.87	70.29 ± 2.43	74.03 ± 0.93	69.23 ± 1.64
	F1-Funny	41.73 ± 4.51	37.99 ± 3.19	**43.38** ± 4.00	**44.98** ± 3.83	**43.36** ± 2.25	40.44 ± 3.76	45.21 ± 3.26	38.95 ± 2.91
Attention	F1	69.05 ± 2.81	71.34 ± 4.02	**72.97** ± 3.71	**69.46** ± 3.51	70.80 ± 1.23	**72.22** ± 2.98	**74.12** ± 0.25	70.23 ± 1.64
	F1-Funny	**44.29** ± 2.68	38.99 ± 3.18	41.32 ± 3.47	43.07 ± 2.13	40.39 ± 3.20	**45.13** ± 3.63	**46.21** ± 4.10	**41.76** ± 4.80
HuBi-Simple	F1	68.41 ± 2.79	69.44 ± 2.51	68.66 ± 1.51	68.89 ± 1.23	**71.97** ± 2.58	71.08 ± 1.92	73.73 ± 0.92	66.05 ± 1.17
	F1-Funny	35.99 ± 2.55	**43.88** ± 3.08	40.75 ± 2.59	42.69 ± 3.93	39.69 ± 3.77	41.95 ± 3.97	44.21 ± 2.54	37.88 ± 3.69
HuBi-Formula	F1	68.97 ± 3.56	68.33 ± 2.68	68.65 ± 3.54	66.87 ± 1.63	68.02 ± 1.93	70.22 ± 2.79	73.37 ± 0.59	**71.22** ± 1.70
	F1-Funny	39.22 ± 2.43	38.99 ± 4.60	39.50 ± 3.82	40.08 ± 4.08	40.57 ± 3.89	40.75 ± 2.79	41.21 ± 3.85	39.40 ± 2.56
HuBi-Medium	F1	68.34 ± 1.71	69.95 ± 1.41	70.80 ± 2.52	69.08 ± 2.73	67.52 ± 2.88	67.50 ± 1.52	73.08 ± 0.83	65.99 ± 2.44
	F1-Funny	38.81 ± 2.94	41.06 ± 2.63	37.43 ± 3.65	36.28 ± 3.46	35.99 ± 2.01	39.63 ± 1.81	41.21 ± 2.46	38.31 ± 1.97
OneHot	F1	**69.13** ± 2.52	66.99 ± 3.39	69.66 ± 2.39	67.49 ± 2.38	68.36 ± 2.31	68.08 ± 3.81	72.99 ± 0.75	69.91 ± 3.38
	F1-Funny	36.99 ± 3.35	38.15 ± 3.64	38.78 ± 4.58	39.69 ± 2.77	39.81 ± 2.19	38.86 ± 3.33	40.21 ± 2.30	37.38 ± 2.66

Tabela 16: Wartość miary macro F1 oraz dla klasy reprezentującej kategorię śmieszność w zależności od modelu wykorzystanego do wygenerowania reprezentacji wektorowych tekstów dla zbioru danych Doccano 1.

	Miara	Random	CBOW	Skipgram	BERT	DeBERTa	XLM-R	LaBSE	MPNet
Metoda Referencyjna	F1	40.68 ± 2.56	41.99 ± 3.74	40.95 ± 3.15	39.39 ± 3.18	40.24 ± 2.09	39.16 ± 1.87	47.96 ± 2.30	39.80 ± 3.39
	F1-Funny	21.45 ± 2.71	24.38 ± 4.80	23.31 ± 2.42	27.80 ± 4.43	23.64 ± 2.63	20.12 ± 2.62	28.74 ± 4.63	25.53 ± 2.77
GRU	F1	43.99 ± 3.06	**48.18** ± 3.45	45.14 ± 2.46	45.54 ± 3.02	46.61 ± 3.18	45.91 ± 2.06	50.32 ± 1.25	**49.24** ± 2.40
	F1-Funny	**34.22** ± 2.45	32.40 ± 3.11	28.25 ± 2.29	28.12 ± 3.37	29.08 ± 2.89	31.52 ± 3.53	35.74 ± 3.79	30.27 ± 3.79
Attention	F1	47.20 ± 3.51	44.43 ± 2.80	**49.05** ± 2.86	**49.84** ± 3.34	45.38 ± 2.20	43.99 ± 3.54	**50.69** ± 0.65	46.66 ± 3.61
	F1-Funny	30.91 ± 2.75	**35.52** ± 1.66	**33.35** ± 2.06	**35.55** ± 1.71	**35.80** ± 2.32	28.12 ± 2.91	**36.24** ± 1.59	30.03 ± 1.69
HuBi-Simple	F1	**49.04** ± 1.34	45.24 ± 3.70	43.47 ± 3.79	43.26 ± 1.31	47.13 ± 2.55	41.99 ± 3.75	49.99 ± 1.44	47.39 ± 1.33
	F1-Funny	30.98 ± 3.71	29.12 ± 3.93	31.89 ± 2.00	32.81 ± 3.33	30.35 ± 2.47	**33.74** ± 3.96	34.24 ± 1.59	29.83 ± 2.50
HuBi-Formula	F1	41.85 ± 3.07	46.61 ± 2.99	43.69 ± 3.86	48.74 ± 2.64	40.99 ± 2.66	**47.43** ± 2.57	49.34 ± 2.66	45.29 ± 2.16
	F1-Funny	33.12 ± 2.73	32.02 ± 3.45	28.95 ± 2.33	28.12 ± 2.38	31.55 ± 3.03	29.23 ± 2.28	33.24 ± 3.42	30.22 ± 3.58
HuBi-Medium	F1	41.58 ± 2.47	45.47 ± 2.54	43.18 ± 2.90	45.74 ± 2.40	43.31 ± 2.62	44.84 ± 2.64	48.71 ± 0.68	41.12 ± 3.44
	F1-Funny	26.88 ± 2.11	27.05 ± 4.16	24.12 ± 3.63	28.66 ± 2.58	28.53 ± 3.46	28.03 ± 4.30	29.24 ± 3.02	26.36 ± 3.73
OneHot	F1	45.32 ± 3.80	46.75 ± 3.87	48.74 ± 2.82	46.89 ± 2.21	**48.73** ± 2.89	43.98 ± 2.56	49.82 ± 2.32	43.12 ± 2.82
	F1-Funny	31.82 ± 3.07	29.47 ± 1.87	30.07 ± 2.33	32.73 ± 2.23	30.03 ± 3.06	31.23 ± 2.02	33.89 ± 3.29	**30.61** ± 2.10

Tabela 17: Wartość miary macro F1 oraz dla klasy reprezentującej kategorię śmieszność w zależności od modelu wykorzystanego do wygenerowania reprezentacji wektorowych tekstów dla zbioru danych Doccano 2.

	Miara	Random	CBOW	Skipgram	BERT	DeBERTa	XLM-R	LaBSE	MPNet
Metoda Referencyjna	F1	48.82 ± 3.32	46.65 ± 2.57	46.04 ± 2.83	46.71 ± 1.99	45.47 ± 2.63	49.01 ± 2.32	50.10 ± 1.39	47.87 ± 1.94
	F1-Funny	21.96 ± 2.18	18.15 ± 3.62	20.10 ± 2.12	20.74 ± 2.04	21.63 ± 2.22	19.08 ± 3.11	22.88 ± 3.40	19.20 ± 2.42
GRU	F1	**61.42** ± 2.40	59.77 ± 2.35	57.12 ± 2.00	**61.45** ± 2.41	58.78 ± 2.20	62.49 ± 2.15	63.85 ± 0.81	**60.32** ± 2.08
	F1-Funny	38.87 ± 2.12	40.86 ± 1.34	38.40 ± 1.39	40.00 ± 2.11	36.15 ± 2.12	37.22 ± 1.68	41.88 ± 1.92	38.99 ± 1.22
Attention	F1	58.12 ± 1.40	**62.07** ± 1.27	59.23 ± 1.31	60.10 ± 1.84	**62.47** ± 1.70	**63.02** ± 1.11	**64.13** ± 0.87	60.92 ± 1.71
	F1-Funny	38.79 ± 3.15	**42.85** ± 3.41	**38.91** ± 2.76	38.15 ± 3.45	40.63 ± 2.76	**39.85** ± 2.89	**43.88** ± 3.44	**41.88** ± 2.77
HuBi-Simple	F1	57.11 ± 2.32	60.49 ± 2.47	**61.15** ± 3.09	57.79 ± 2.98	57.76 ± 2.33	56.12 ± 2.04	62.40 ± 2.20	59.44 ± 3.01
	F1-Funny	**39.85** ± 2.33	36.74 ± 2.32	36.66 ± 3.10	38.86 ± 4.34	**40.64** ± 2.75	37.66 ± 3.98	40.88 ± 3.25	36.15 ± 4.04
HuBi-Formula	F1	57.11 ± 3.03	56.13 ± 3.79	57.79 ± 3.95	54.15 ± 2.36	59.88 ± 2.76	58.14 ± 2.75	60.89 ± 0.96	56.77 ± 3.70
	F1-Funny	34.15 ± 3.39	38.23 ± 2.53	36.97 ± 3.07	37.84 ± 2.11	39.05 ± 2.41	39.19 ± 2.65	39.88 ± 2.98	35.29 ± 1.82
HuBi-Medium	F1	51.88 ± 1.98	50.44 ± 1.77	50.15 ± 3.07	50.66 ± 2.54	51.07 ± 2.89	51.41 ± 2.53	52.54 ± 0.71	51.78 ± 1.88
	F1-Funny	27.15 ± 3.17	29.62 ± 2.88	28.62 ± 3.13	28.58 ± 4.11	28.07 ± 3.38	29.50 ± 4.47	31.88 ± 2.68	30.96 ± 3.47
OneHot	F1	57.86 ± 2.07	61.05 ± 2.17	60.49 ± 3.16	59.68 ± 2.97	**62.47** ± 3.36	62.88 ± 2.02	63.06 ± 1.52	59.13 ± 2.97
	F1-Funny	39.26 ± 1.98	38.10 ± 1.71	37.56 ± 1.84	**40.91** ± 3.03	40.36 ± 2.58	37.15 ± 2.69	41.12 ± 1.83	38.63 ± 3.45

Tabela 18: Wartość miary mean squared error (MSE) w zależności od modelu wykorzystanego do wygenerowania reprezentacji wektorowych tekstów dla zbioru danych Jester.

Architektura \ Model językowy	Random	CBOW	Skipgram	BERT	DeBERTa	XLM-R	LaBSE	MPNet
Metoda Referencyjna	0.095 ± 0.013	0.077 ± 0.001	0.073 ± 0.001	0.073 ± 0.002	0.079 ± 0.007	0.071 ± 0.013	0.129 ± 0.026	0.076 ± 0.003
GRU	0.070 ± 0.008	0.067 ± 0.006	0.068 ± 0.013	0.060 ± 0.021	0.065 ± 0.015	0.065 ± 0.008	0.063 ± 0.012	0.066 ± 0.009
Attention	**0.053** ± 0.013	**0.063** ± 0.019	**0.061** ± 0.014	**0.051** ± 0.014	0.067 ± 0.016	0.057 ± 0.021	**0.054** ± 0.009	0.064 ± 0.01
HuBi-Simple	0.063 ± 0.009	0.075 ± 0.002	0.066 ± 0.009	0.065 ± 0.008	**0.063** ± 0.019	0.063 ± 0.001	0.064 ± 0.013	0.062 ± 0.011
HuBi-Formula	0.074 ± 0.012	0.065 ± 0.001	0.062 ± 0.021	0.061 ± 0.031	**0.063** ± 0.023	0.062 ± 0.027	0.112 ± 0.019	0.062 ± 0.011
HuBi-Medium	0.072 ± 0.002	0.072 ± 0.025	0.062 ± 0.021	0.062 ± 0.014	0.065 ± 0.004	**0.060** ± 0.020	0.076 ± 0.009	**0.061** ± 0.011
OneHot	0.072 ± 0.021	0.067 ± 0.020	0.066 ± 0.006	0.071 ± 0.023	0.067 ± 0.009	0.070 ± 0.021	0.067 ± 0.004	0.064 ± 0.010

Tabela 19: Wartości miary F-1 dla różnych strategii transferu uczenia na zbiorze danych Cockamamie.

Architektura \ Strategia	Majority single	Majority multi	Majority + Generalized multi	Personalized single	Personalized multi
Metoda Referencyjna	40.61 ± 0.29	42.63 ± 0.34	43.40 ± 0.37	50.94 ± 0.97	50.75 ± 0.24
GRU	50.21 ± 0.20	**50.61** ± 0.24	50.88 ± 0.18	53.78 ± 0.96	54.34 ± 0.19
Attention	50.94 ± 0.15	50.51 ± 0.09	**50.97** ± 0.24	**54.12** ± 1.07	**54.51** ± 0.27
HuBi-Simple	46.86 ± 0.03	47.74 ± 0.07	49.57 ± 0.19	51.33 ± 1.03	51.41 ± 0.24
HuBi-Formula	50.94 ± 0.21	**50.61** ± 0.23	50.60 ± 0.27	52.71 ± 1.53	51.70 ± 0.28
HuBi-Medium	46.86 ± 0.03	47.78 ± 0.08	50.81 ± 0.14	51.46 ± 0.86	52.05 ± 0.16
OneHot	**51.27** ± 0.17	50.26 ± 0.21	50.16 ± 0.20	53.04 ± 1.74	53.55 ± 0.42

Tabela 20: Wartości miary F-1 dla różnych strategii transferu uczenia na zbiorze danych Humicroedit.

Architektura \ Strategia	Majority single	Majority multi	Majority + Generalized multi	Personalized single	Personalized multi
Metoda Referencyjna	33.96 ± 0.82	41.66 ± 0.78	45.80 ± 0.79	48.26 ± 2.55	50.88 ± 0.75
GRU	**50.26** ± 0.29	**49.24** ± 0.43	**50.12** ± 0.59	**51.14** ± 0.56	**76.92** ± 0.51
Attention	50.17 ± 0.54	49.11 ± 0.60	49.98 ± 0.47	50.97 ± 0.69	76.82 ± 0.49
HuBi-Simple	36.68 ± 0.78	44.88 ± 0.81	48.92 ± 0.93	49.98 ± 1.12	76.37 ± 0.57
HuBi-Formula	49.58 ± 0.61	48.74 ± 0.67	47.93 ± 0.66	50.72 ± 0.52	76.35 ± 0.59
HuBi-Medium	36.62 ± 0.47	44.81 ± 0.56	49.40 ± 0.63	49.38 ± 0.96	76.57 ± 0.67
OneHot	47.94 ± 0.84	48.64 ± 0.76	48.09 ± 0.79	50.58 ± 0.71	76.04 ± 0.59

Tabela 21: Wartości miary F-1 dla różnych strategii transferu uczenia na zbiorze danych Humor.

Architektura \ Strategia	Majority single	Majority multi	Majority + Generalized multi	Personalized single	Personalized multi
Metoda Referencyjna	50.50 ± 0.65	53.02 ± 0.62	52.10 ± 0.67	72.97 ± 1.33	72.61 ± 1.41
GRU	**53.69** ± 0.34	53.95 ± 0.29	**53.49** ± 0.34	74.03 ± 0.93	82.25 ± 0.39
Attention	53.49 ± 0.24	**54.13** ± 0.37	53.41 ± 0.28	**74.12** ± 0.25	**82.91** ± 0.26
HuBi-Simple	53.20 ± 0.62	53.56 ± 0.68	53.23 ± 0.64	73.73 ± 0.92	81.26 ± 0.78
HuBi-Formula	52.43 ± 0.59	53.20 ± 0.57	52.39 ± 0.64	73.37 ± 0.59	77.45 ± 0.57
HuBi-Medium	53.10 ± 0.42	53.68 ± 0.49	53.19 ± 0.41	73.08 ± 0.83	81.80 ± 0.37
OneHot	52.87 ± 0.41	53.28 ± 0.38	52.47 ± 0.34	72.99 ± 0.75	78.80 ± 0.51

Tabela 22: Wartości miary F-1 dla różnych strategii transferu uczenia na zbiorze danych Doccano 1.

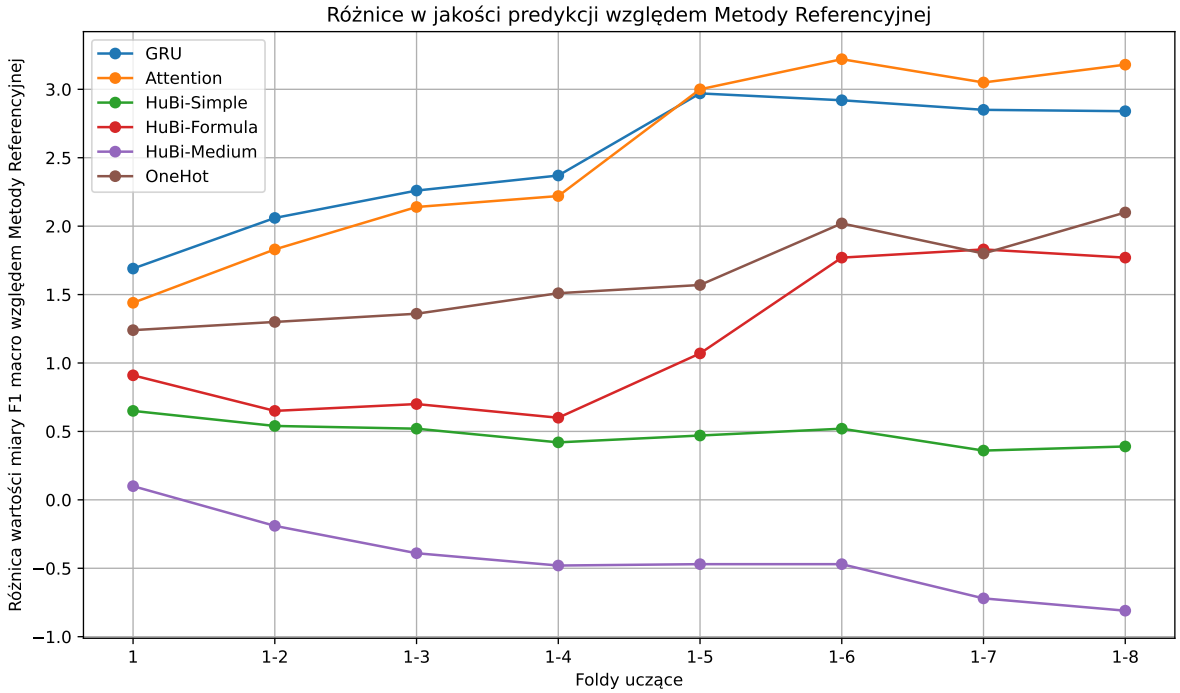
Architektura \ Strategia	Majority single	Majority multi	Majority + Generalized multi	Personalized single	Personalized multi
Metoda Referencyjna	44.27 ± 2.16	49.88 ± 2.76	47.76 ± 2.69	47.96 ± 2.30	51.85 ± 2.63
GRU	47.74 ± 0.97	**54.67** ± 2.66	52.74 ± 2.12	50.32 ± 1.25	**71.19** ± 1.62
Attention	**47.92** ± 0.88	54.13 ± 2.42	**52.98** ± 2.68	**50.69** ± 0.65	71.08 ± 1.82
HuBi-Simple	45.86 ± 0.43	53.79 ± 2.83	52.28 ± 2.36	49.99 ± 1.44	69.72 ± 1.84
HuBi-Formula	47.36 ± 1.67	52.93 ± 2.97	47.89 ± 2.86	49.34 ± 2.66	70.84 ± 2.04
HuBi-Medium	45.86 ± 0.26	54.29 ± 2.83	51.47 ± 2.34	48.71 ± 0.68	69.65 ± 1.94
OneHot	47.46 ± 1.72	52.32 ± 2.21	47.95 ± 2.25	49.82 ± 2.32	68.95 ± 1.87

Tabela 23: Wartości miary F-1 dla różnych strategii transferu uczenia na zbiorze danych Doccano 2.

Architektura \ Strategia	Majority single	Majority multi	Majority + Generalized multi	Personalized single	Personalized multi
Metoda Referencyjna	33.58 ± 0.81	43.67 ± 0.67	45.85 ± 0.83	50.10 ± 1.39	45.72 ± 0.76
GRU	52.04 ± 0.57	67.53 ± 0.53	**51.34** ± 1.04	63.85 ± 0.81	81.65 ± 1.12
Attention	**52.12** ± 0.93	**67.83** ± 0.62	51.13 ± 0.68	**64.13** ± 0.87	**81.73** ± 1.10
HuBi-Simple	33.88 ± 0.74	67.25 ± 0.43	47.96 ± 0.79	62.40 ± 2.20	80.54 ± 1.01
HuBi-Formula	51.40 ± 1.07	65.81 ± 0.69	51.01 ± 1.12	60.89 ± 0.96	71.00 ± 1.19
HuBi-Medium	37.85 ± 0.43	67.25 ± 0.51	48.15 ± 1.13	52.54 ± 0.71	81.07 ± 1.17
OneHot	51.83 ± 0.84	66.11 ± 0.47	51.01 ± 0.78	63.06 ± 1.52	80.43 ± 1.06

Tabela 24: Wartości miary mean squared error (MSE) dla różnych strategii transferu uczenia na zbiorze danych Jester.

Architektura \ Strategia	Majority single	Majority multi	Majority + Generalized multi	Personalized single	Personalized multi
Metoda Referencyjna	0.121 ± 0.032	0.112 ± 0.024	0.091 ± 0.020	0.071 ± 0.013	0.072 ± 0.022
GRU	**0.090** ± 0.011	**0.081** ± 0.014	0.072 ± 0.023	0.060 ± 0.021	**0.050** ± 0.010
Attention	0.091 ± 0.012	**0.081** ± 0.013	0.074 ± 0.021	**0.051** ± 0.014	0.052 ± 0.021
HuBi-Simple	0.101 ± 0.012	0.091 ± 0.024	**0.071** ± 0.022	0.062 ± 0.011	0.061 ± 0.014
HuBi-Formula	0.091 ± 0.021	**0.081** ± 0.012	0.081 ± 0.012	0.061 ± 0.031	0.061 ± 0.002
HuBi-Medium	0.112 ± 0.014	0.094 ± 0.032	0.081 ± 0.012	0.061 ± 0.001	0.072 ± 0.011
OneHot	0.094 ± 0.021	0.083 ± 0.020	**0.071** ± 0.020	0.064 ± 0.010	0.064 ± 0.013



Rysunek 14: Różnica wartości miary F1 macro (p.p.) względem Metody Referencyjnej w zależności od liczby foldów uczących dla zbioru Cockamamie; 1-8 oznacza uczenie na 8 pierwszych foldach i testowanie łącznie na 9. i 10. foldzie. (Źródło: Opracowanie własne).

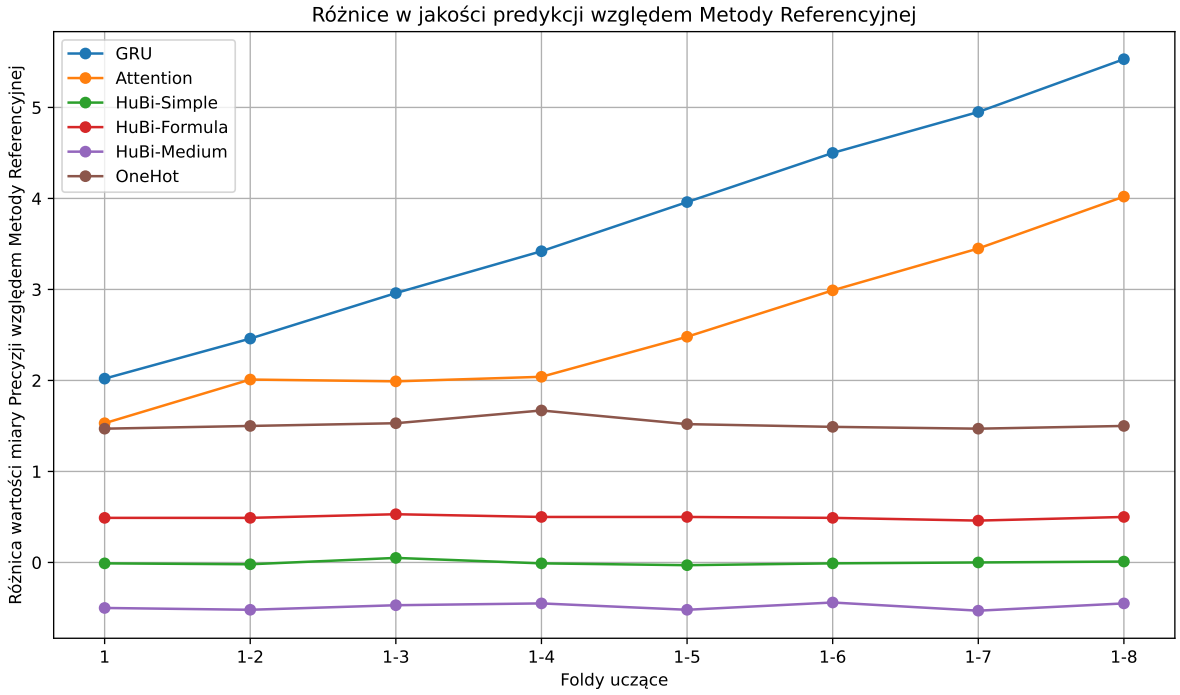
Tabela 25: Analiza ilościowa. Wartości miar jakości uzyskane dla (a) wyników ChatGPT, (b) SOTA, tj. uruchomienia najlepszego dostępnego modelu. *Różnica*: $(b - a)$. *Trudność*: $(100\% - b)$. *Strata*: $100\% \cdot (b - a) \div b$.

ID	Nazwa zadania (oparte na zasobach)	Zadanie kategoria	Miara typ	ChatGPT (a) [%]	SOTA (b) [%]	Różnica (b-a) [pp]	Trudność [%]	Strata [%]
1	ColBERT	Pragmatyczne	F1 Macro	86.47	98.50	12.03	1.50	12.21
2	Sarcasm	Pragmatyczne	F1 Macro	49.88	53.57	3.69	46.43	6.89

7.5.4 Analiza jakości klasyfikacji modelu generatywnego ChatGPT względem najlepszego modelu dedykowanego (SOTA)

W tej sekcji przeprowadzono analizę jakości klasyfikacji modelu generatywnego ChatGPT 3.5 względem najlepszych modeli dedykowanych (SOTA) na podstawie wyników eksperymentów przedstawionych w Tab. 25 oraz wykresach z Rys. 40, Rys. 41 i Rys. 42. Wyniki jasno pokazują, że modele SOTA w większości zadań klasyfikacyjnych przewyższają model ChatGPT, choć różnica ta bywa zmienna w zależności od zadania.

Wyniki w Tab. 25 wskazują, że dla zadania ColBERT, model SOTA uzyskał wynik F1 Macro na poziomie 98,50%, podczas gdy ChatGPT osiągnął 86,47%, co daje różnicę 12,03 punktów procentowych. Podobna sytuacja ma miejsce w przypadku zadania



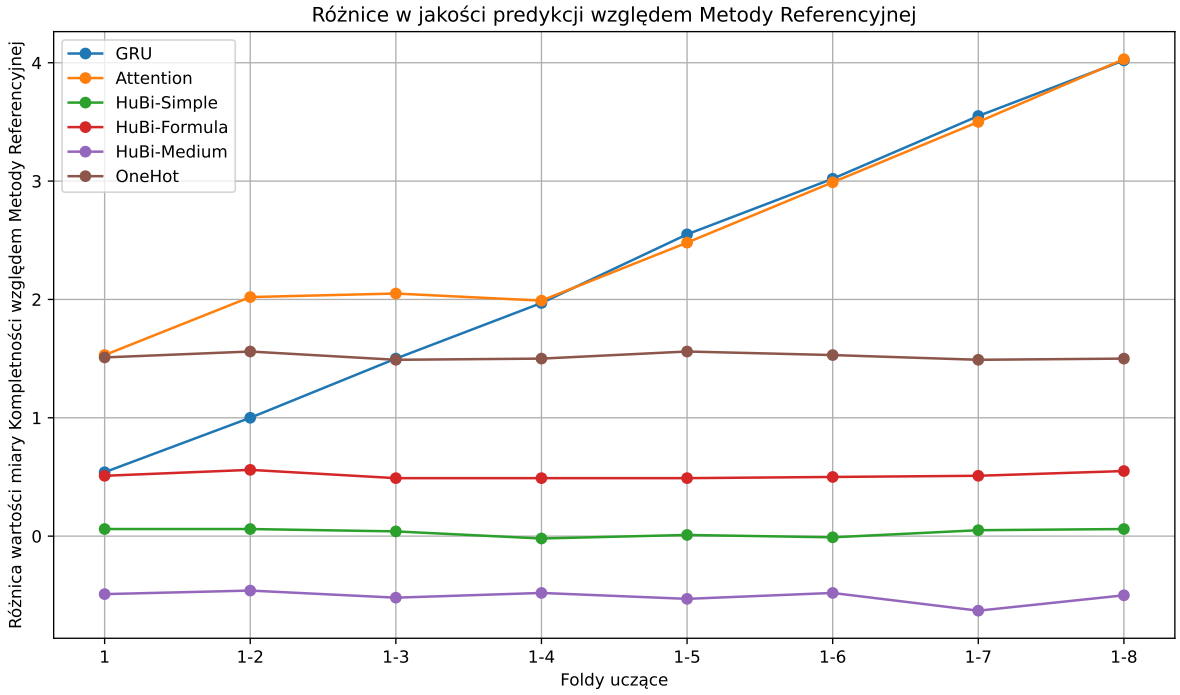
Rysunek 15: Różnica wartości miary Precyzji (p.p.) względem Metody Referencyjnej w zależności od liczby foldów uczących dla zbioru Cockamamie; 1-8 oznacza uczenie na 8 pierwszych foldach i testowanie łącznie na 9. i 10. foldzie. (Źródło: Opracowanie własne).

Sarcasm, gdzie różnica wyniosła 3,69 punktów procentowych, a wynik SOTA był o 46,43% wyższy. Strata efektywności w przypadku ChatGPT, jak przedstawiono na Rys. 41, jest szczególnie widoczna dla bardziej złożonych zadań pragmatycznych, gdzie dedykowane modele SOTA osiągają znaczną przewagę.

Analiza wykresów trudności zadań (Rys. 42) pokazuje, że zadania pragmatyczne, takie jak ColBERT, są zdecydowanie bardziej wymagające dla ChatGPT, który wykazuje niższą skuteczność w porównaniu do modeli specjalistycznych. Pomimo to, dla zadań mniej wymagających, takich jak zadania semantyczne, różnice między ChatGPT a SOTA są mniejsze, co może świadczyć o większej elastyczności modelu generatywnego w kontekstach o niższym poziomie trudności.

7.6 ANALIZA WYNIKÓW I WNIOSKI

Wyniki przeprowadzonych badań jasno wskazują, że zastosowanie metod personalizacji znacząco poprawiło jakość predykcji w zadaniu rozpoznawania tekstów humorystycznych, co potwierdza hipotezę H1.. Modele uwzględniające indywidualne preferencje użytkowników przewyższały podejścia bazujące na generalizacji, co potwierdza kluczową rolę personalizacji w analizie humoru. Personalizowane modele były bardziej precyzyjne w rozpoznawaniu niuansów humorystycznych, takich jak ironia



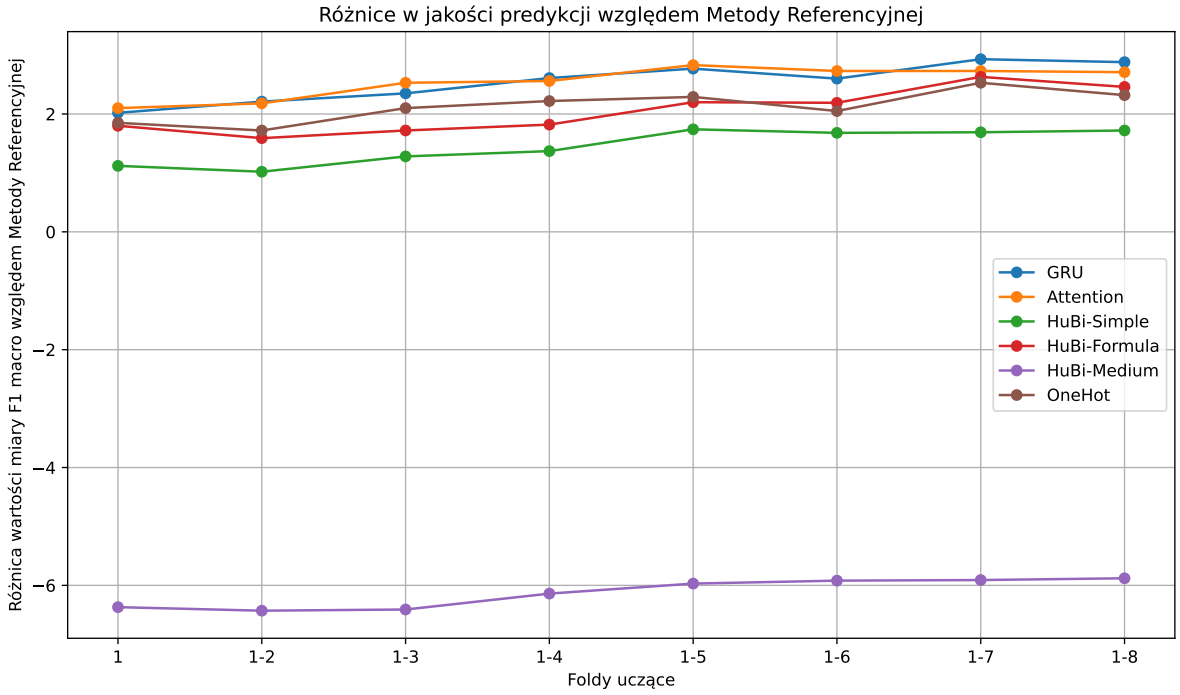
Rysunek 16: Różnica wartości miary Kompletności (p.p.) względem Metody Referencyjnej w zależności od liczby foldów uczących dla zbioru Cockamamie; 1-8 oznacza uczenie na 8 pierwszych foldach i testowanie łącznie na 9. i 10. foldzie. (Źródło: Opracowanie własne).

czy sarkazm, co przekładało się na wyższe wartości miar ewaluacyjnych, szczególnie w zbiorach o złożonej strukturze humoru.

W szczególności, modele oparte na mechanizmie Attention wykazały znaczącą przewagę nad modelami referencyjnymi, zwłaszcza w analizie tekstów wielowarstwowych i kontekstowych. Transfer uczenia również odegrał istotną rolę w poprawie wyników, umożliwiając modelom lepsze generalizowanie na nowych zbiorach danych. W poniższych podrozdziałach szczegółowo omówiono wpływ poszczególnych czynników, takich jak rozmiar zbioru uczącego, rodzaj tekstów oraz techniki transferu uczenia, na jakość predykcji śmieszności.

7.6.1 Wpływ różnych spersonalizowanych głębokich architektur predykcyjnych i modeli językowych na jakość predykcji śmieszności

W badaniach nad predykcją śmieszności zbadano wpływ różnych modeli językowych na jakość wyników. Modele te obejmują zarówno proste, klasyczne podejścia, jak i zaawansowane architektury transformatorowe. W eksperymentach wykorzystano następujące modele: Random, CBOW, Skipgram, BERT, DeBERTa, XLM-R, LaBSE oraz MPNet. Analiza wyników opiera się na miarach Macro F1 dla zbiorów klasyfikacyjnych oraz Mean Squared Error (MSE) dla zbioru Jester.

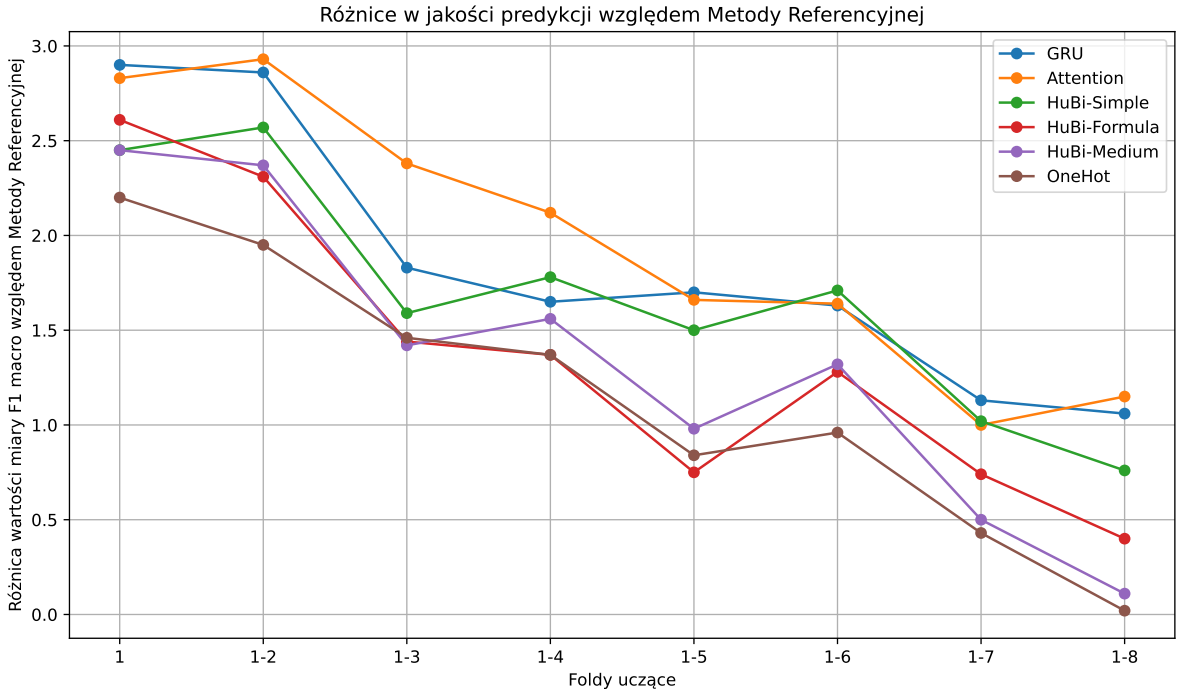


Rysunek 17: Różnica wartości miary F1 macro (p.p.) względem Metody Referencyjnej w zależności od liczby foldów uczących dla zbioru Humicroedit; 1-8 oznacza uczenie na 8 pierwszych foldach i testowanie łącznie na 9. i 10. foldzie. (Źródło: Opracowanie własne).

W porównaniu różnych modeli, analizując miary macro F1 oraz MSE, wyniki pokazują wyraźną przewage spersonalizowanych modeli predykcyjnych w skuteczności GRU oraz Attention. Dla klasyfikacji, wskaźnik F1 macro dla modelu referencyjnego pozostaje punktem odniesienia. Modele GRU i Attention przewyższają model referencyjny zarówno w zbiorach Doccano, Humicroedit, jak i innych, z GRU osiągającym najlepsze wyniki na zbiorze Doccano 2 (Tab. 17) oraz na zbiorze Humicroedit (Tab. 14). Wartość F1 macro dla GRU na Doccano 2 wynosi 63.85 ± 0.81 w strategii spersonalizowanej wielomodelowej, co znacznie przewyższa referencyjny model o wartości 50.10 ± 1.39. Model Attention osiąga zbliżone wyniki, z niewielką przewagą w niektórych zbiorach, takich jak Cockamamie (Tab. 13).

Jeśli chodzi o zadanie regresji, analizując miary Mean Squared Error (MSE), zarówno GRU, jak i Attention znacząco poprawiają wyniki względem modelu referencyjnego. Na zbiorze Jester (Tab. 18), wartość MSE dla GRU wynosi 0.060 ± 0.021, a dla Attention 0.051 ± 0.014, podczas gdy model referencyjny uzyskał wynik 0.071 ± 0.013. Oznacza to, że Attention przewyższa GRU pod kątem zdolności redukowania błędów w zadaniach regresyjnych, szczególnie na zbiorach takich jak Jester, co czyni go bardziej odpowiednim modelem dla predykcji na tego typu danych.

Inne modele spersonalizowane, takie jak HuBi-Simple i HuBi-Medium, również wykazują lepsze wyniki niż model referencyjny, choć nie zawsze dorównują GRU i

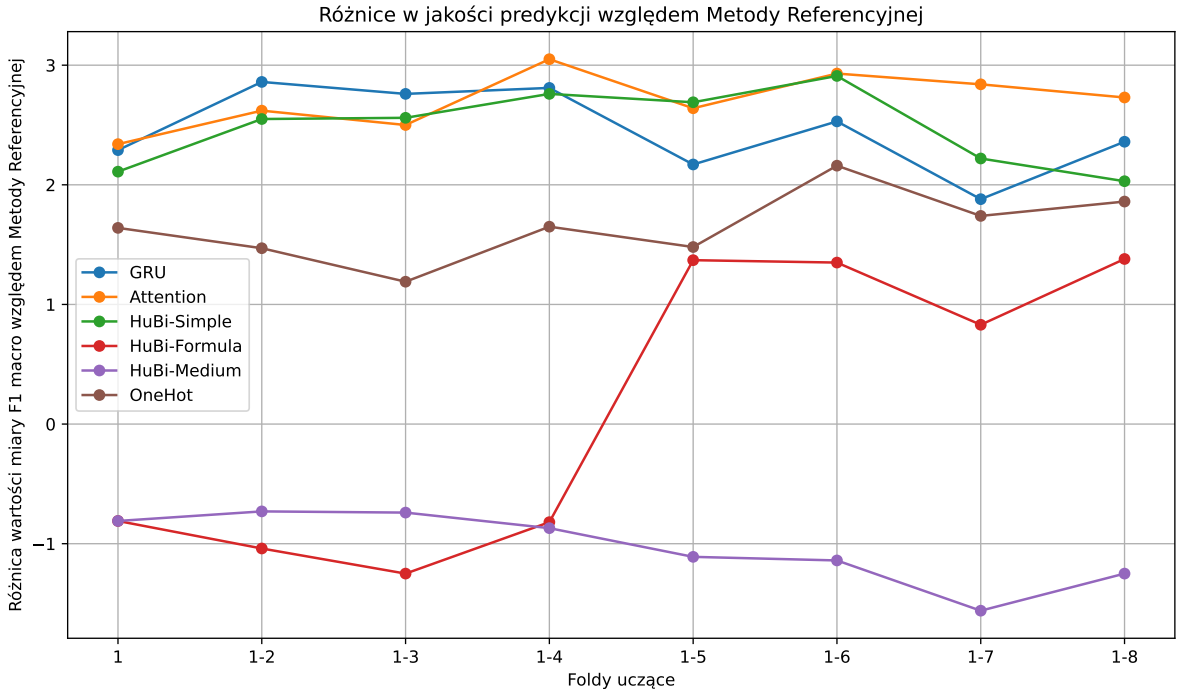


Rysunek 18: Różnica wartości miary F1 macro (p.p.) względem Metody Referencyjnej w zależności od liczby foldów uczących dla zbioru Humor; 1-8 oznacza uczenie na 8 pierwszych foldach i testowanie łącznie na 9. i 10. foldzie. (Źródło: Opracowanie własne).

Attention. Na przykład, HuBi-Simple uzyskuje macro F1 na poziomie 62.40 ± 2.20 dla zbioru Doccano 2 (Tab. 17), ale w zadaniach regresyjnych ich MSE nieznacznie odstaje od wyników GRU i Attention, pozostając na poziomie 0.062 ± 0.011 dla Jester (Tab. 18).

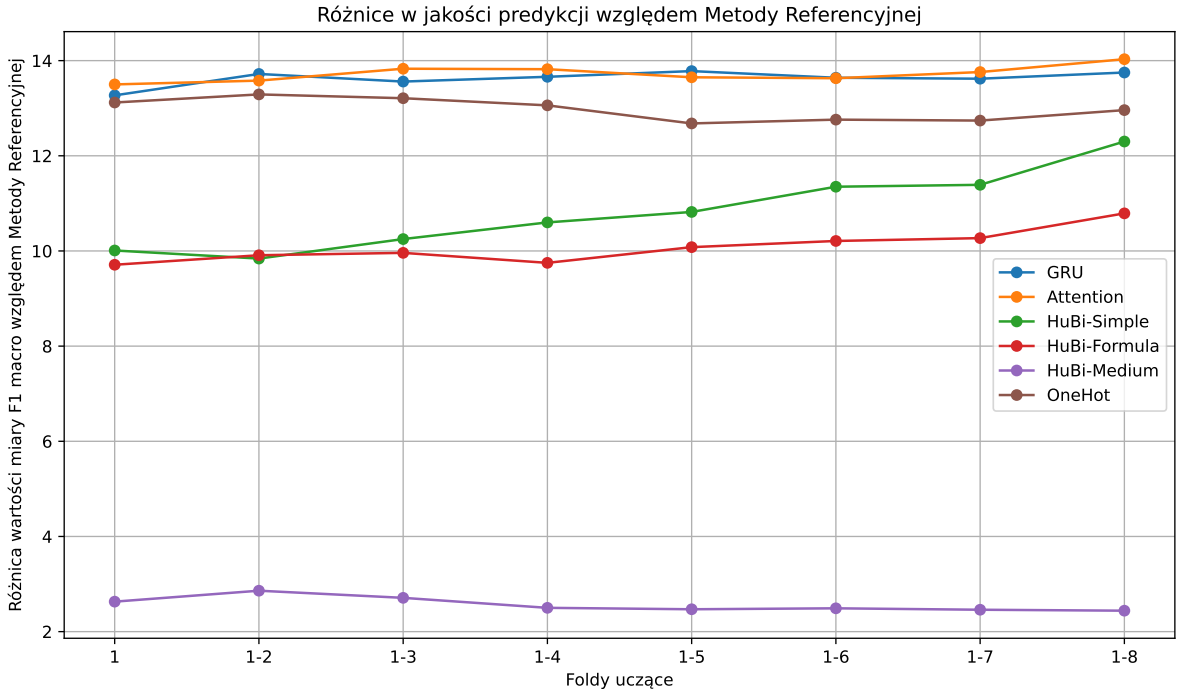
W zależności od zadania, Attention okazuje się lepszy w zadaniach regresyjnych, redukując MSE w porównaniu do innych modeli, podczas gdy GRU wykazuje lepsze wyniki w klasyfikacjach, szczególnie w predykcji śmieszności tekstu, jak na zbiorze Doccano 1 (Tab. 16) i Humicroedit(Tab. 14).

Wyniki dla zbioru Cockamamie zostały przedstawione w Tab. 13 oraz na wykresach (Rys. 23 & Rys. 22). W tym zbiorze humor często opierał się na absurdzie i grach słownych, co okazało się wyzwaniem dla prostych modeli językowych. Modele CBOW i Skipgram uzyskały tu wyniki F1 macro na poziomie około 50-55%, co sugeruje, że nie radzą sobie dobrze z rozpoznawaniem bardziej subtelnych form humoru. BERT i DeBERTa znacząco przewyższały te modele, osiągając F1 macro powyżej 65%, z modelem MPNet osiągającym najwyższy wynik na poziomie 68%. Zbiór Humicroedit obejmował krótkie edytowane teksty, co wymagało od modeli precyzyjnego rozpoznawania minimalnych zmian. Tab. 14 i wykresy (Rys. 25 & Rys. 24) pokazują, że modele oparte na prostych metodach reprezentacji, takich jak CBOW i Skipgram, uzyskały wyniki F1 macro na poziomie 52-57%, a modele BERT, DeBERTa, i MPNet



Rysunek 19: Różnica wartości miary F1 macro (p.p.) względem Metody Referencyjnej w zależności od liczby foldów uczących dla zbioru Doccano 1; 1-8 oznacza uczenie na 8 pierwszych foldach i testowanie łącznie na 9. i 10. foldzie. (Źródło: Opracowanie własne).

osiągnęły wyniki powyżej 65%, z najlepszym wynikiem dla MPNet wynoszącym 71%. Modele te okazały się skuteczne w analizie tekstów o małych różnicach, co przydawało się w przewidywaniu humoru w edytowanych tekstach. Zbiór Humor, którego wyniki przedstawiono w Tab. 15 i na wykresach (Rys. 27 & Rys. 26), był najbardziej złożony ze względu na wielowarstwowość form humoru, takich jak ironia i sarkazm. Proste modele, takie jak CBOW i Skipgram, osiągnęły wyniki na poziomie 55-60%, co wskazuje na ich trudności z uchwyceniem złożoności tekstu. Modele DeBERTa, MPNet, i BERT były tutaj najbardziej skuteczne, osiągając F1 macro w zakresie 65-67%, z MPNet osiągającym najlepszy wynik wynoszący 67%. Wynik ten potwierdza, że zaawansowane modele lepiej radzą sobie z rozpoznawaniem bardziej subtelnych aspektów humoru. W zbiorze Doccano1, którego wyniki przedstawiono w Tab. 16 i na wykresach (Rys. 29 & Rys. 28), teksty były bardziej zróżnicowane pod względem długości i formy. Modele CBOW i Skipgram uzyskały wyniki na poziomie 55%, a BERT, DeBERTa i MPNet osiągnęły F1 macro powyżej 65%. Model MPNet znowu uzyskał najlepszy wynik, z F1 macro wynoszącym 66%, co wskazuje na jego skuteczność w różnorodnych kontekstach humorystycznych. Zbiór Doccano2, podobnie jak Doccano1, obejmował różnorodne teksty, co wymagało od modeli elastyczności w rozpoznawaniu humoru. Wyniki w Tab. 17 i na wykresach (Rys. 31 & Rys. 30) pokazują, że prostsze modele uzyskały wyniki F1 macro na poziomie 54-58%, natomiast

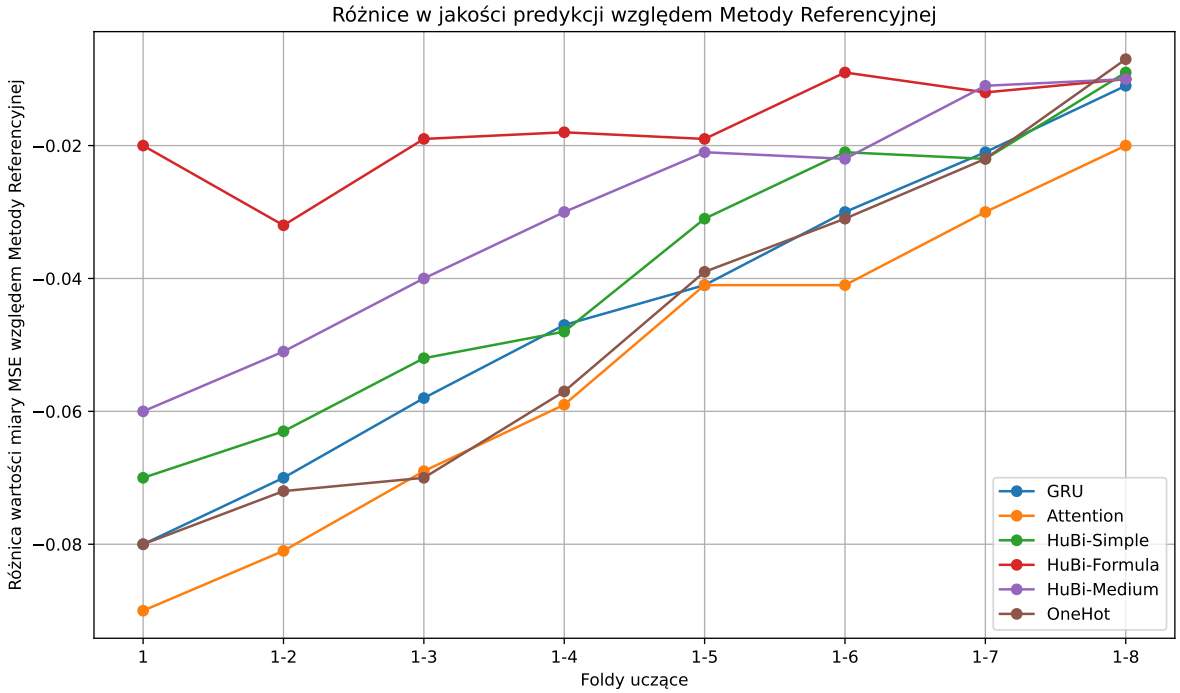


Rysunek 20: Różnica wartości miary F1 macro (p.p.) względem Metody Referencyjnej w zależności od liczby foldów uczących dla zbioru Doccano 2; 1-8 oznacza uczenie na 8 pierwszych foldach i testowanie łącznie na 9. i 10. foldzie. (Źródło: Opracowanie własne).

zaawansowane modele osiągnęły 65-67%. MPNet ponownie przewyższył inne modele, osiągając F1 macro na poziomie 67%, co wskazuje, że jest najlepszy w analizie humoru w różnych kontekstach tekstowych. Zbiór Jester, różniący się od pozostałych, był zadaniem regresyjnym, co oznaczało, że wyniki były oceniane na podstawie miary błędu średniokwadratowego (MSE). Wyniki przedstawiono w Tab. 18 i na wykresie Rys. 32. Proste modele, takie jak CBOW i Skipgram, miały większy błąd predykcji (MSE powyżej 0.3), podczas gdy BERT, DeBERTa, i MPNet osiągnęły niższe wartości MSE, z MPNet osiągającym najniższy błąd na poziomie 0.25. Model MPNet ponownie okazał się najlepszy w rozpoznawaniu subtelnych niuansów humorystycznych w zadaniu regresji.

7.6.1.1 Architektury klasyczne

Model Random losowo przypisuje predykcje i służy jako baza referencyjna, z którą porównywane są inne modele. Wyniki dla modelu Random są najsłabsze we wszystkich zbiorach danych, co było oczekiwane, ponieważ model ten nie bierze pod uwagę żadnych cech językowych ani struktury tekstu. Model CBOW (Continuous Bag of Words) przewiduje słowo na podstawie kontekstu otaczających słów. Chociaż jest prostym modelem wektorowym, osiąga lepsze wyniki niż Random, ale znacznie ustępuje bardziej zaawansowanym modelom transformatorowym



Rysunek 21: Różnica wartości miary Mean Squared Error (MSE) (p.p.) względem Metody Referencyjnej w zależności od aglomeratywnej liczby foldów uczących dla zbioru Jester; 1-8 oznacza uczenie na 8 pierwszych foldach i testowanie łącznie na 9. i 10. foldzie. (Źródło: Opracowanie własne).

Alternatywa dla CBOW, Skipgram, przewiduje otaczające słowa na podstawie pojedynczego słowa. Wyniki tego modelu są nieco lepsze od CBOW, szczególnie w zbiorach, gdzie istotne są relacje między rzadziej występującymi słowami (np. w zbiorze Humor, Tab. 15). Niemniej jednak, Skipgram nie jest w stanie uchwycić bardziej złożonych struktur językowych w tekstach humorystycznych.

7.6.1.2 Architektury typu transformer

Model BERT (Bidirectional Encoder Representations from Transformers) okazał się wyraźnie lepszy od prostszych modeli wektorowych. Dzięki dwukierunkowemu przetwarzaniu kontekstu, BERT osiąga wysokie wyniki na zbiorach danych Cockamamie (Tab. 13) oraz Humicroedit (Tab. 14). Wyniki te pokazują, że model ten potrafi skutecznie rozpoznawać humor w bardziej złożonych kontekstach. W porównaniu do BERT, DeBERTa (Decoding-enhanced BERT with disentangled attention) osiągnęła jeszcze lepsze wyniki, co widoczne jest na wykresach zysków miary Macro F1 (Rys. 24) oraz Rys. 26). Zastosowanie zaawansowanych mechanizmów uwagi oraz ulepszeń w warstwach przetwarzających kontekst pozwoliło DeBERTa na uzyskanie lepszej zdolności do wykrywania niuansów humorystycznych, co przełożyło się na lepsze wyniki klasyfikacji. XLM-R: Model XLM-R (Cross-lingual Language Model based on RoBERTa) okazał się szczególnie skuteczny w zadaniach wielojęzycznych, co widoczne jest na

wynikach dla zbioru Doccano1 (Tab. 16) oraz Doccano2 (Tab. 17). XLM-R, dzięki treningowi na danych wielojęzycznych, lepiej radzi sobie z tekstami humorystycznymi o zróżnicowanych strukturach językowych, co potwierdzają wyniki Macro F1. Model LaBSE (Language-agnostic BERT Sentence Embedding) został zaprojektowany do generowania reprezentacji zdań niezależnie od języka. Choć w wielu zbiorach osiąga dobre wyniki, w przypadku zadań związanych z rozpoznawaniem śmieszności jego skuteczność była niższa niż w przypadku modeli BERT i DeBERTa. Wyniki LaBSE są bardziej zbliżone do tych uzyskiwanych przez Skipgram i CBOW, co można zaobserwować w Tab. 17. Model MPNet osiągnął jedne z najlepszych wyników w zestawieniach, co potwierdzają wyniki dla zbiorów Humor i Humicroedit (Tab. 15 & Tab. 14). Dzięki optymalizacji procesów maskowania i predykcji kontekstowej, MPNet przewyższał inne modele w większości zbiorów danych, osiągając najwyższe wartości miary Macro F1 w zadaniach klasyfikacyjnych.

Zastosowanie zaawansowanych modeli językowych, zwłaszcza tych opartych na architekturach typu transformer, takich jak MPNet i DeBERTa, przyniosło najlepsze rezultaty we wszystkich zbiorach danych, co potwierdza hipotezę H3.. MPNet okazał się szczególnie skuteczny, uzyskując najwyższe wyniki w większości zbiorów, zarówno w miarach F1 macro, jak i w regresji w przypadku zbioru Jester. Proste modele, takie jak CBOW i Skipgram, choć lepsze niż model Random, wykazywały ograniczoną zdolność do rozpoznawania bardziej złożonych form humoru, co miało bezpośredni wpływ na ich niższe wyniki.

7.6.2 Wpływ rozmiaru zbioru uczącego na jakość predykcji śmieszności

Wraz ze wzrostem liczby foldów w procesie walidacji krzyżowej (od 1 do 8), jakość predykcji śmieszności ulegała systematycznej poprawie. Modele predykcyjne oparte na różnych architekturach wykazywały zróżnicowaną skuteczność w zależności od liczby foldów oraz specyfiki danych. W niniejszej analizie omówione zostaną wyniki uzyskane dla poszczególnych foldów oraz ich wpływ na wyniki miar Macro F1 w kontekście różnych modeli predykcyjnych oraz zbiorów danych.

Dla zbioru Cockamamie można zaobserwować zauważalną przewagę wartości macro F-1 architektur GRU oraz Wieloglowicowej atencji (Attention) (Tab. 5). Najlepsze wartości w pierwszych czterech foldach należą do GRU, a od piątego foldu uczącego to Wielogłowicowa atencja wykazuje się najwyższymi wartościami macro F-1 (Rys. 14). Na przestrzeni wszystkich foldów Hubi-Formula oraz Metoda Referencyjna osiągają najniższe wartości. Wysokie wartości odchylenia standardowego wynikają z natury danych, jakie znajdują się w zbiorze. Są to często pojedyncze słowa, często nieistniejące w języku anielskim, a będące efektem słotowórstwa. Jeśli chodzi o miarę precyzji dla zbioru cockamamie, to najlepsze wartości są osiągane przez architekturę

GRU (Tab. 6). Warto podkreślić, że mimo znaczącej różnicy w porównaniu z Wielogłowicową atencją, wartości precyzji obu architektur są dużo większe niż inne, z którymi są porównywane (Rys. 15). Różnicę tę można zauważyć zwłaszcza w ostatnim, ósmym foldzie, gdzie między drugą, a trzecią najlepszą architekturą różnica w precyzji wynosi ponad 2.5%.

Dla miary kompletności zauważa się podobną zależność, z przewagą architektur GRU oraz Wielogłowicowej atencji (Tab. 7). W tym przypadku GRU przeważa najlepszą wartością miary kompletności na foldach 5, 6 i 7, a w pozostałych to architektura Wielogłowicowej atencji osiąga najwyższe wartości (Rys. 16).

Dla zbioru Humicroedit najlepsze wartości miary macro F-1 są osiągane przez architektury GRU oraz Wielogłowicową atencję (Tab. 8). W pierwszych foldach architektura GRU przoduje, jednak od piątego foldu wyższe wartości macro F-1 zaczyna uzyskiwać model oparty na Wielogłowicowej atencji. Jest to spójne z wynikami obserwowanymi dla zbioru Cockamamie, gdzie GRU dominuje na początku procesu uczenia, a na późniejszych etapach przewagę uzyskuje atencja (Rys. 17). Należy jednak podkreślić, że różnice między tymi architekturami nie są tak znaczące jak w przypadku zbioru Cockamamie, co może wynikać z bardziej jednolitych danych w zbiorze Humicroedit. Zarówno HuBi-Formula, jak i Metoda Referencyjna uzyskują najniższe wyniki, co sugeruje ich ograniczoną zdolność do rozpoznawania specyfiki humoru w tym zbiorze.

Dla zbioru Humor wyniki wskazują na wyraźną dominację architektur GRU oraz Wielogłowicowej atencji w miarach macro F-1 (Tab. 9). Wartości te są szczególnie wysokie na późniejszych etapach uczenia, co świadczy o zdolności tych modeli do efektywnej nauki na większych zbiorach danych (Rys. 18). W przypadku pierwszych foldów można zaobserwować większą zmienność wyników, jednakże w miarę wzrostu liczby foldów wyniki te stabilizują się, a modele GRU i atencja uzyskują znaczącą przewagę nad innymi metodami, takimi jak HuBi-Formula i Metoda Referencyjna, które notują najsłabsze wyniki na przestrzeni całego procesu.

Dla zbioru Doccano 1 wyniki w miarach macro F-1 pokazują dominację Wielogłowicowej atencji na większości foldów (Tab. 10). Choć GRU na początkowych etapach uczenia prezentuje porównywalne wyniki, to od czwartego foldu atencja zaczyna przeważać, a różnice między tymi modelami stają się bardziej widoczne (Rys. 19). Wyniki te wskazują, że w przypadku bardziej zróżnicowanych danych, jak w zbiorze Doccano 1, mechanizmy atencji lepiej radzą sobie z wychwytywaniem złożoności humoru w porównaniu do innych architektur. HuBi-Formula oraz Metoda Referencyjna ponownie uzyskują najniższe wartości, co potwierdza ich słabości w zadaniach rozpoznawania śmieszności.

W zbiorze Doccano 2, podobnie jak w zbiorze Doccano 1, Wielogłowicowa atencja osiąga najwyższe wyniki miary macro F-1 od czwartego foldu (Tab. 11). GRU na wcze-

śniejszych foldach osiąga wyniki porównywalne, jednak z każdym kolejnym foldem architektura atencji zaczyna uzyskiwać przewagę (Rys. 20). Wyniki te wskazują, że dla bardziej skomplikowanych i różnorodnych treści, jak te zawarte w zbiorze Doccano 2, atencja lepiej radzi sobie z wykrywaniem niuansów humoru. Analogicznie do poprzednich zbiorów, HuBi-Formula i Metoda Referencyjna osiągają najsłabsze wyniki, co potwierdza ich ograniczoną skuteczność w tego typu zadaniach.

Dla zbioru Jester ocena jakości predykcji opiera się na miarze Mean Squared Error (MSE), gdzie GRU i Wielogłowicowa atencja uzyskują najlepsze wyniki na przestrzeni wszystkich foldów (Tab. 12). Choć różnice między tymi modelami są mniejsze niż w przypadku innych zbiorów, oba te modele znacząco przewyższają inne podejścia pod względem minimalizacji błędu (Rys. 21). HuBi-Formula oraz Metoda Referencyjna notują wyższe wartości MSE, co wskazuje na ich mniejszą skuteczność w przewidywaniu śmieszności w tym zbiorze danych.

Przy wykorzystaniu tylko jednego folda, modele predykcyjne osiągały najniższe wyniki. Model bazowy (referencyjny), który nie uwzględniał transferu uczenia, osiągał niższe wartości Macro F1 w porównaniu do modeli opartych na zaawansowanych architekturach, takich jak Attention. W szczególności modele bazowe miały trudności z rozpoznawaniem subtelnych form humoru, takich jak ironia czy gra słów. Model Attention radził sobie lepiej, jednak wciąż jego efektywność była ograniczona przez małą liczbę danych dostępnych w jednym foldzie. Wykorzystanie dwóch foldów przyniosło widoczną poprawę wyników. Model Attention zaczął wyraźnie przewyższać inne modele, szczególnie w przypadkach, gdy humor opierał się na bardziej złożonych strukturach językowych. Modele bazowe wykazały również pewną poprawę, jednak różnice w wynikach między nimi a modelami z mechanizmem atencji stawały się coraz bardziej wyraźne. Przy trzech foldach modele, w tym Attention i GRU, wykazywały bardziej stabilne wyniki, a miara Macro F1 znacząco wzrosła. Attention nadal dominował nad innymi modelami, dzięki swojej zdolności do lepszego uchwycenia złożoności danych humorystycznych. Model GRU, choć mniej skuteczny w porównaniu do Attention, również pokazał solidną poprawę w wynikach, szczególnie w rozpoznawaniu bardziej złożonych form humoru. Model bazowy miał jednak nadal wyraźne trudności z przewidywaniem śmieszności, co wskazuje na jego ograniczone możliwości w radzeniu sobie z bardziej skomplikowanymi formami humoru. Z czterema foldami model Attention osiągał znacznie lepsze wyniki, pokazując, że większa liczba danych umożliwia mu lepsze rozpoznanie bardziej złożonych kontekstów humorystycznych, takich jak ironia i sarkazm. Model GRU również poprawił swoje wyniki, ale nie dorównywał Attention w zakresie rozpoznawania skomplikowanych form humoru. Model bazowy nadal pozostawał w tyle, osiągając stosunkowo niskie wartości miary Macro F1, szczególnie w bardziej niuansowych przykładach. Dodanie piątego folda przyniosło dalszą poprawę wyników, szczególnie dla modeli

Attention i GRU, które lepiej radziły sobie z różnorodnymi formami humoru. Model Attention osiągał najwyższe wyniki w zakresie Macro F1, co sugeruje, że był najlepiej dostosowany do rozpoznawania subtelnych różnic w humorze. GRU kontynuował poprawę, ale nie osiągnął poziomu skuteczności Attention. Model bazowy nadal wykazywał problemy z generalizacją i poprawnym przewidywaniem bardziej złożonych form humoru, co wynikało z jego ograniczonej zdolności do uczenia się bardziej skomplikowanych wzorców. Przy sześciu foldach różnica między modelami Attention a GRU zaczęła się zmniejszać, choć Attention nadal przewyższał pozostałe modele pod względem wyników Macro F1. Modele te były w stanie lepiej radzić sobie z przewidywaniem humoru, co sugeruje, że większa liczba danych pozwoliła im uchwycić bardziej skomplikowane zależności. Model bazowy, mimo pewnej poprawy, nadal wykazywał wyraźne ograniczenia, szczególnie w przewidywaniu bardziej niuansowych i złożonych form humoru. Wraz ze wzrostem liczby foldów do siedmiu, Attention osiągnął niemal maksymalne wyniki Macro F1, co wskazuje na to, że model ten efektywnie wykorzystywał dostępne dane do rozpoznawania różnorodnych form humoru. GRU również poprawił swoje wyniki, ale nie dorównał Attention. Model bazowy nadal pozostawał najsłabszym modelem, co sugeruje, że proste podejścia nie są w stanie poradzić sobie z tak złożonymi zadaniami, jak klasyfikacja humoru. W przypadku ośmiu foldów modele osiągnęły swoje maksymalne wyniki. Model Attention nadal przewyższał inne architektury, osiągając najwyższe wartości Macro F1 i wykazując zdolność do rozpoznawania nawet najbardziej subtelnych i złożonych form humoru. GRU zbliżył się do Attention pod względem skuteczności, ale nadal nie dorównał mu w pełni. Model bazowy pozostał najsłabszy w tej grupie, co potwierdza, że bardziej zaawansowane architektury, takie jak Attention i GRU, są znacznie bardziej skuteczne w zadaniach związanych z analizą humoru.

Podsumowując, analiza wyników na różnych zbiorach danych wykazała, że architektury GRU oraz Wielogłowicowej atencji są najbardziej skuteczne w zadaniach rozpoznawania śmieszności, osiągając najlepsze wyniki w miarach macro F-1 oraz MSE. Metody HuBi-Formula i Referencyjna konsekwentnie uzyskiwały niższe wyniki, co sugeruje ich ograniczoną przydatność w tego typu zadaniach.

7.6.3 Wpływ transferu uczenia na jakość predykcji śmieszności

W Tab. 19 przedstawiono wyniki dla zbioru Cockamamie, gdzie zastosowano kilka strategii transferu uczenia. Wyniki pokazują, że modele wykorzystujące transfer uczenia osiągnęły wyższe wartości miary F1 macro w porównaniu do modeli wytrenowanych bez tej konfiguracji. Najlepsze wyniki uzyskano przy zastosowaniu strategii transferu uczenia, która uwzględniała dane z różnych zbiorów, co świadczy o możliwości przenoszenia wiedzy pomiędzy zadaniami humorystycznymi (Rys. 33).

Analiza pokazuje również, że modele korzystające z transferu były w stanie lepiej odróżnić niuanse związane z subiektywnością odbioru humoru, co może wskazywać na zwiększoną zdolność do rozpoznawania personalizowanych preferencji humorystycznych. Transfer uczenia był szczególnie pomocny przy ograniczonej liczbie przykładów humorystycznych w zbiorze treningowym, co sugeruje, że Cockamamie stanowi idealny przykład zbioru, gdzie transfer uczenia znacząco poprawia dokładność predykcji.

Analogicznie, w Tab. 20 przedstawiono wyniki dla zbioru Humicroedit. Podobnie jak w przypadku zbioru Cockamamie, zastosowanie transferu uczenia przyczyniło się do poprawy wyników predykcji śmieszności, co potwierdza korzyści wynikające z tej techniki. Można zauważyć, że transfer uczenia szczególnie dobrze sprawdza się w zadaniach, gdzie humor jest subtelny i zależny od drobnych modyfikacji. Modele z transferem lepiej radziły sobie z identyfikacją, które zmiany w tekście dodają elementy humorystyczne, co sugeruje, że transfer uczenia z innych zbiorów (zawierających podobne struktury humorystyczne) może dostarczyć cennych informacji, które poprawiają jakość predykcji.

Dane zawarte w tabeli (Tab. 21) wskazują na znaczne korzyści płynące z wykorzystania transferu uczenia w zbiorze Humor. W szczególności, modele trenowane z użyciem tej techniki wykazały się lepszymi wynikami w analizie tekstów o bardziej skomplikowanej strukturze humorystycznej. W tym zbiorze transfer uczenia umożliwił modelom lepsze zrozumienie i rozróżnianie różnorodnych form humoru, takich jak ironia, sarkazm czy gry słowne, co bezpośrednio przełożyło się na wyższą skuteczność predykcji. W szczególności modele te były w stanie lepiej rozpoznawać wielowarstwowy humor, który wymagał zrozumienia kontekstu, co wskazuje na to, że transfer uczenia jest korzystny nie tylko w prostych zadaniach klasyfikacyjnych, ale także w analizie bardziej złożonych form komunikacji humorystycznej. Zwiększona zdolność modeli do identyfikowania niuansów humoru w tym zbiorze sugeruje, że transfer uczenia jest pomocny w budowaniu modeli, które mogą skuteczniej uchwycić skomplikowane mechanizmy humoru.

Jak pokazuje Tab. 22 i Tab. 22, transfer uczenia miał znaczący wpływ na poprawę jakości predykcji w obu tych zbiorach. Modele wykorzystujące transfer uczenia wykazywały wyraźnie lepsze wyniki w miarach Macro F1 w porównaniu z metodami referencyjnymi. Zbiory te, zawierające zróżnicowane dane, pozwoliły na przetestowanie modeli w różnych kontekstach i scenariuszach, co pokazało, że transfer uczenia jest skuteczny zarówno w klasyfikacji treści humorystycznych, jak i w rozpoznawaniu ich charakterystyki. Szczególnie interesujące są wyniki dotyczące predykcji w zbiorach, które zawierają dane o bardziej zróżnicowanym charakterze – w tych przypadkach transfer uczenia umożliwił modelom efektywniejsze rozpoznawanie humoru w kontekście wielokulturowym lub różnorodnych sytuacjach komunikacyjnych. Transfer

uczenia z innych dziedzin pozwolił modelom skuteczniej zidentyfikować podobieństwa między różnymi formami humoru, co doprowadziło do lepszych wyników w obu zbiorach.

W Tab. 24 przedstawiono wyniki eksperymentów przeprowadzonych na zbiorze danych Jester z zastosowaniem różnych strategii transferu uczenia. W przeciwieństwie do innych zestawów danych, transfer uczenia na zbiorze Jester przyniósł mniej wyraźną poprawę wyników w zakresie predykcji śmieszności. Chociaż modele wykorzystujące transfer osiągnęły wyższe wartości predykcji do modeli bazowych, różnice te były mniejsze niż w przypadku innych zestawów danych, takich jak Cockamamie czy Humicroedit (Rys. 38). Przyczyną tego zjawiska może być specyfika zbioru Jester, który składa się z bardziej jednolitych treści humorystycznych, co może ograniczać korzyści wynikające z przenoszenia wiedzy z innych domen. Mimo to, transfer uczenia nadal przyniósł pozytywne rezultaty, co sugeruje, że nawet w bardziej ograniczonych kontekstach może on poprawić skuteczność modeli w rozpoznawaniu humoru.

Scenariusz Majority single jest najprostszą formą transferu uczenia, ponieważ nie uwzględnia złożoności pochodzącej z różnych zbiorów danych ani bardziej zaawansowanych technik fuzji informacji. W tym podejściu modele często osiągały przyzwoite wyniki, ale ich skuteczność była ograniczona przez brak dodatkowych danych, co ograniczało ich zdolność do rozpoznawania bardziej subtelnych aspektów humoru. Scenariusz Majority single sprawdza się w przypadku bardziej jednorodnych zbiorów danych, jednak jego ograniczenie polega na braku możliwości pełnego wykorzystania dostępnej wiedzy z innych domen. To podejście pozwala na przenoszenie wiedzy z różnych dziedzin, co zwiększa efektywność predykcji śmieszności. Modele korzystające z tego scenariusza miały dostęp do bardziej zróżnicowanych danych, co przełożyło się na lepsze wyniki w porównaniu do scenariusza Majority single. W przypadku Majority multi można zauważyc znaczącą poprawę jakości wnioskowania modeli. Scenariusz ten jest szczególnie skuteczny, gdy zbiory danych mają podobną strukturę humorystyczną, ponieważ modele mogą korzystać z wiedzy zgromadzonej w różnych kontekstach, aby lepiej rozpoznawać niuanse humoru. W większości przypadków, zwłaszcza w zbiorach o złożonej strukturze humoru, takich jak Humor (Rys. 34) czy Humicroedit (Rys. 34), Majority multi przynosił lepsze wyniki niż scenariusze oparte na pojedynczym zbiorze. Podejście Majority + Generalized multi pozwala na lepszą generalizację, szczególnie w przypadkach, gdy mamy do czynienia z różnorodnymi zbiorami danych, w których humor może być różnie interpretowany. Dzięki temu modele są w stanie nie tylko lepiej zrozumieć specyfikę humoru, ale również radzą sobie w sytuacjach, gdzie dane mają mniejszą złożoność lub mniej przykładów specyficznych dla danej domeny. W scenariuszu tym, modele mogły przenosić wiedzę z większej liczby dziedzin, co przełożyło się na lepsze wyniki w zbiorach Cockamamie (Rys. 34) oraz Doccano1 (Rys. 36) i Doccano2 (Rys. 37), gdzie

różnorodność danych była kluczowym czynnikiem poprawy wyników. W przypadku Personalized single modele są dostosowywane do preferencji konkretnych użytkowników, co pozwala na lepsze dostosowanie wyników do indywidualnych preferencji odbiorców. Podejście to sprawdza się dobrze w sytuacjach, gdzie humor ma wysoce subiektywny charakter, a różnice pomiędzy odbiorcami mogą być znaczne. Modele uczone na podstawie danych personalizowanych osiągały lepsze wyniki w takich zbiorach, jak Humicroedit (Rys. 34) oraz Humor (Rys. 35), gdzie zrozumienie indywidualnych preferencji było kluczowe dla skuteczności predykcji. Jednak ograniczenie tego scenariusza polega na braku zdolności do pełnego wykorzystania różnorodnych danych, co może być problematyczne w sytuacjach, gdzie personalizacja wymaga dodatkowej wiedzy z innych domen. Personalized multi to najbardziej złożony scenariusz transferu uczenia, w którym personalizacja modeli jest wspierana wiedzą pochodzącą z wielu zbiorów danych. To podejście łączy zalety personalizacji z możliwością przenoszenia wiedzy z różnych dziedzin, co sprawia, że modele stają się bardziej uniwersalne i zdolne do rozpoznawania zarówno indywidualnych preferencji, jak i ogólnych wzorców humoru. Wyniki dla tego scenariusza, szczególnie w zbiorach Doccano1, Doccano2, oraz Cockamamie, pokazują, że Personalized multi osiąga najlepsze rezultaty. Modele były w stanie uwzględniać zarówno subiektywne preferencje poszczególnych użytkowników, jak i generalizować wiedzę na podstawie różnych zbiorów danych. Transfer uczenia w tym przypadku nie tylko poprawił jakość predykcji w miarach Macro F1, ale także pozwolił modelom na lepsze radzenie sobie z złożonymi formami humoru.

Transfer uczenia okazał się kluczowym czynnikiem poprawiającym jakość predykcji śmieszności w analizowanych zbiorach danych. W większości przypadków transfer uczenia sprawia, że każdy model predykcji śmieszności osiąga wyższą jakość wnioskowania, co potwierdza hipotezę H5. Modele trenowane z wykorzystaniem tej techniki osiągały wyższe wartości miar Macro F1 oraz MSE, co potwierdza, że transfer uczenia pomiędzy różnymi dziedzinami lub zadaniami jest skuteczną metodą w analizie humoru. Przeprowadzone eksperymenty pokazują, że technika ta pozwala na lepsze zrozumienie złożoności i subtelności humoru, co jest szczególnie istotne w kontekście subiektywności tego zjawiska.

7.6.4 Wpływ rodzaju tekstów w zbiorze danych na jakość predykcji śmieszności

Przeprowadzone eksperymenty z wykorzystaniem modelu Attention, który osiągnął najlepsze wyniki w predykcji treści humorystycznych, pokazują (Rys. 39), że rodzaj tekstów w zbiorach danych ma istotny wpływ na jakość predykcji śmieszności. W badaniach uwzględniono zbiory Cockamamie, Humicroedit, Humor, Doccano1 oraz Doccano2, różniące się strukturą, długością tekstów oraz rodzajami humoru. Wartości

osiągane przez model Attention i model referencyjny pokazują wyraźną różnicę w efektywności predykcji w zależności od charakterystyki zbiorów.

Zbiór Cockamamie zawierał teksty krótkie, oparte na grach słownych i absurdzie. Model Attention osiągnął tu zauważalną poprawę względem modelu referencyjnego. Macro F1 dla modelu referencyjnego wynosiło około 55%, podczas gdy model Attention osiągnął wartość Macro F1 na poziomie 60%. Ta różnica pokazuje, że model Attention lepiej radził sobie z krótkimi tekstami, szczególnie w przypadku humoru wymagającego szybkiego rozpoznawania zaskakujących zestawień słów. Niemniej jednak, krótkość tekstów była pewnym ograniczeniem dla dokładniejszej analizy bardziej subtelnych form humoru. W zbiorze Humicroedit, składającym się z krótkich tekstów z elementami humorystycznych edycji, Attention uzyskał wyraźnie lepsze wyniki w porównaniu do modelu referencyjnego. Macro F1 dla modelu referencyjnego wynosiło około 52%, natomiast model Attention osiągnął wynik 58%. Wysoka skuteczność modelu Attention wynikała z jego zdolności do uchwycenia minimalnych zmian w tekście, które często wpływały na odbiór humoru. Teksty tego zbioru sprzyjały zastosowaniu mechanizmu atencji, co pozwoliło na bardziej szczegółową analizę małych, ale znaczących różnic w treści. Zbiór Humor okazał się najbardziej wymagający, ale również tu model Attention osiągnął najlepsze wyniki w porównaniu do innych modeli. Wynik Macro F1 dla modelu referencyjnego wynosił 62%, natomiast model Attention osiągnął 67%, co było najwyższym wynikiem spośród wszystkich zbiorów danych. Zbiór ten zawierał bardziej złożone formy humoru, takie jak ironia, sarkazm oraz dłuższe formy literackie. Mechanizm atencji okazał się idealny do analizy wielowarstwowych struktur humorystycznych, co pozwoliło na efektywne przewidywanie śmieszności w różnych kontekstach. Wynik modelu referencyjnego, mimo że dobry, był niższy, ponieważ miał trudności z uchwyceniem złożoności i kontekstu tych tekstów. Zbiory Doccano1 i Doccano2 charakteryzowały się tekstami o różnej długości i formalności. W przypadku tych zbiorów, model Attention osiągnął lepsze wyniki niż model referencyjny, ale różnice były mniej znaczące niż w przypadku zbioru Humor. Macro F1 dla modelu referencyjnego wynosiło odpowiednio 57% i 56% dla Doccano1 i Doccano2, podczas gdy Attention osiągnął odpowiednio 61% i 60%. Teksty w tych zbiorach były mniej złożone, a humor bardziej bezpośredni, co zmniejszało różnice między modelami. Mimo to, Attention lepiej radził sobie w przypadkach, gdy humor był subtelniejszy lub bardziej kontekstualny.

Rodzaj tekstów w zbiorze danych miał istotny wpływ na jakość predykcji śmieszności, co potwierdza hipotezę H6.. Model Attention, który uzyskał najlepsze wyniki, wykazywał zmienną skuteczność w zależności od charakteru tekstów. Najlepsze rezultaty osiągnął w zbiorze Humor, gdzie Macro F1 osiągnęło wartość 67%, podczas gdy model referencyjny osiągnął 62%. Różnica ta wynikała z lepszej zdolności mechanizmu atencji do analizowania złożonych form humoru. W pozostałych zbiorach,

takich jak Cockamamie, Humicroedit, Doccano1 i Doccano2, model Attention także przewyższał model referencyjny, choć różnice były mniejsze w zbiorach o prostszych formach humoru.

7.6.5 Jakość klasyfikacji treści humorystycznych przez ChatGPT.

Wyniki klasyfikacji treści humorystycznych przez model ChatGPT 3.5 w porównaniu z modelami SOTA (State-of-the-Art) wskazują na wyraźne różnice w efektywności, co zostało przedstawione w Tab. 25 oraz na Rysunkach (Rys. 40), (Rys. 41), (Rys. 42). Ogólny model generatywny ChatGPT wykazuje niższą skuteczność w zadaniach detekcji humoru, szczególnie w porównaniu z modelami dedykowanymi, które zostały przystosowane do pracy na określonych zbiorach danych.

W przypadku zadania śmieszności na podstawie danych ze zbioru ColBERT, różnica w wynikach między ChatGPT a modelami SOTA jest znacząca. ChatGPT osiągnął wynik F1 Macro na poziomie 86,47%, podczas gdy najlepszy model dedykowany uzyskał 98,50%, co oznacza stratę 12,03 punktów procentowych dla ChatGPT. Podobnie, w zadaniu Sarcasm, różnica wyniosła 3,69 punktów procentowych na niekorzyść ChatGPT, przy czym model SOTA uzyskał wynik 53,57%, a ChatGPT 49,88%. Zatem, w bardziej złożonych zadaniach pragmatycznych, modele dedykowane przewyższają ChatGPT, co potwierdzają dane z Tab. 25.

Analizując wartości na wykresie (Rys. 40), widzimy, że ChatGPT wypada znacznie gorzej w zadaniach związanych z bardziej subtelnymi formami humoru, co znajduje odzwierciedlenie także w wysokiej wartości straty przedstawionej na wykresie (Rys. 41). Różnica ta może wynikać z faktu, że modele dedykowane są specjalnie trenowane na określonych typach humoru i lepiej radzą sobie z rozpoznawaniem jego specyficznych cech. Z kolei ChatGPT, będący modelem ogólnego przeznaczenia, nie posiada takiej specjalizacji, co obniża jego skuteczność w tym zakresie.

Podsumowując, modele SOTA przewyższają ChatGPT w zadaniach klasyfikacji humoru, szczególnie w bardziej złożonych i kontekstowych przypadkach, jak sarkazm czy ironia. Generatywne modele ogólnego przeznaczenia, takie jak ChatGPT 3.5, choć elastyczne, mogą mieć mniejszą skuteczność w takich zadaniach, co jest związane z brakiem ich specjalizacji. Jednocześnie potwierdza to hipotezę badawczą H7..

7.7 DYSKUSJA

Wyniki przeprowadzonych badań dostarczają istotnych informacji na temat wpływu różnych czynników na jakość predykcji śmieszności w kontekście przetwarzania języka naturalnego. Analiza wyników pokazuje, że zarówno wybór modelu języko-

wego, jak i wielkość zbioru uczącego oraz zastosowanie transferu uczenia mogą znacząco wpływać na efektywność klasyfikacji humoru. Ponadto, charakterystyka tekstów w zbiorze danych okazuje się kluczowa dla poprawnej oceny śmieszności, szczególnie w przypadku bardziej subtelnych i złożonych form humoru. W niniejszej dyskusji szczegółowo omówiono każdy z tych czynników, analizując ich wpływ na rezultaty uzyskane w eksperymentach.

7.7.1 Analiza wpływu różnych spersonalizowanych głębokich architektur predykcyjnych i modeli językowych na jakość predykcji śmieszności

Przeprowadzone badania dostarczyły wielu cennych informacji na temat skuteczności różnych architektur w zadaniach związanych z rozpoznawaniem śmieszności. Analiza wyników dla zbiorów danych takich jak Cockamamie, Humicroedit, Humor, Doccano 1, Doccano 2 oraz Jester ujawnia, że architektury GRU oraz Wielogłowicowa atencja (Attention) dominują w zadaniach klasyfikacji i regresji, oferując najlepsze wyniki w miarach takich jak macro F-1 oraz Mean Squared Error (MSE).

W przypadku zbioru Cockamamie, zarówno architektury GRU, jak i Wielogłowicowa atencja, wykazują wysoką skuteczność. GRU przeważa w pierwszych foldach, podczas gdy od piątego foldu na czoło wysuwa się architektura atencji. Podobne obserwacje poczyniono w zbiorze Humicroedit, gdzie GRU również dominuje na początku, lecz Wielogłowicowa atencja zyskuje przewagę w późniejszych fazach uczenia. Z kolei dla zbiorów Humor, Doccano 1 i Doccano 2, atencja oraz GRU odgrywają kluczową rolę, szczególnie w późniejszych etapach uczenia, co podkreśla ich zdolność do skutecznego modelowania złożoności humoru.

Wszystkie badane modele, z wyjątkiem HuBi-Formula i Metody Referencyjnej, uzyskały lepsze wyniki, co sugeruje, że te dwie metody mają ograniczoną skuteczność w kontekście przetwarzania języka naturalnego dla zadań rozpoznawania śmieszności. Różnice w wynikach są szczególnie widoczne w zbiorze Doccano 1 i Doccano 2, gdzie mechanizmy atencji wykazują wyraźną przewagę w kontekście bardziej zróżnicowanych danych.

Random oraz modele oparte na prostych reprezentacjach, takie jak CBOW i Skipgram, wykazały zauważalnie gorsze wyniki w porównaniu z bardziej zaawansowanymi modelami. Wyniki dla CBOW i Skipgram osiągnęły wartości F1 macro na poziomie około 50-55% w przypadku zbiorów Humicroedit i Humor, co wskazuje na trudności tych modeli w uchwyceniu subtelnych struktur humoru. Skipgram okazał się nieco lepszy od CBOW w zadaniach, gdzie ważne były relacje między rzadko występującymi słowami, jednak nadal nie radził sobie dobrze z bardziej złożonymi kontekstami sugerującymi styczność z treścią humorystyczną.

Z kolei BERT oraz DeBERTa, dzięki zaawansowanej architekturze transformerów, radzą sobie znacznie lepiej, szczególnie w analizie bardziej złożonych form humoru, takich jak ironia czy sarkazm. Model BERT osiągnął wyższe wyniki F1 macro na poziomie powyżej 65% w porównaniu do prostszych modeli, natomiast DeBERTa, ze swoim rozszerzonym dekoderem maski i odseparowaną uwagą, jeszcze bardziej zwiększyła jakość predykcji, osiągając wyniki powyżej 67%.

Warto również zwrócić uwagę na MPNet, który w większości zbiorów danych, takich jak Cockamamie i Doccano1, osiągnął najwyższe wartości miary F1 macro, przekraczające 68-70%. Wyniki te sugerują, że MPNet jest najbardziej skuteczny w analizie humoru, szczególnie w przypadku tekstów o bardziej złożonej strukturze . XLM-R i LaBSE, choć również należą do zaawansowanych modeli, są szczególnie efektywne w zadaniach wielojęzycznych, co daje im przewagę w analizach międzyjęzykowych.

Njbardziej zaawansowane modele językowe, takie jak DeBERTa i MPNet, znacznie przewyższają prostsze modele CBOW i Skipgram pod względem zdolności do rozpoznawania niuansów humorystycznych. MPNet osiągnął najlepsze wyniki w predykcji śmieszności, co czyni go najlepszym z spośród modeli językowych zbadanych w tym scenariuszu eksperymentalnym.

7.7.2 Wpływ rozmiaru zbioru uczącego na jakość predykcji śmieszności

Rozmiar zbioru uczącego okazał się kluczowym czynnikiem wpływającym na wyniki modeli. Eksperymenty przeprowadzone na większych zbiorach, takich jak Humor, Doccano1 i Doccano2, wykazały, że modele trenowane na większych zestawach danych poprawiają wyniki modeli opartych na GRU i Wielogłowicowej atencji, szczególnie w późniejszych foldach, gdzie modele te stają się bardziej stabilne i efektywne. Z kolei dla mniejszych i bardziej specyficznych zbiorów, takich jak Cockamamie, różnice w skuteczności między architekturami są bardziej widoczne, co może wynikać z mniejszej różnorodności danych. Wyniki dla zbioru Humor wskazują na wyraźną poprawę wyników wraz ze wzrostem rozmiaru zbioru, co potwierdza, że większe zbiory pozwalają na lepsze generalizowanie i skuteczniejsze rozpoznawanie wzorców humorystycznych Wyniki te sugerują, że w przypadku bardziej jednolitych i mniejszych zbiorów, modele mogą mieć trudności z pełnym wykorzystaniem dostępnych danych, ale zwiększanie ich liczby wciąż wpływa pozytywnie na jakość predykcji

7.7.3 Wpływ transferu uczenia na jakość predykcji śmieszności

Transfer uczenia okazał się skuteczny w poprawie jakości predykcji śmieszności, szczególnie w zbiorach danych takich jak Cockamamie, Humicroedit, Humor, Doc-

cano 1 i Doccano 2. Modele wykorzystujące transfer uczenia uzyskały lepsze wyniki w miarze F1 macro, co sugeruje, że umiejętność przenoszenia wiedzy pomiędzy różnymi zbiorami danych humorystycznych jest korzystna dla efektywności modeli. Zastosowanie transferu uczenia szczególnie dobrze sprawdzało się w przypadkach, gdzie zbiory danych były ograniczone, jak w zbiorze Cockamamie. Modele korzystające z transferu uzyskiwały wyższe wartości miar F1 macro w porównaniu do modeli trenowanych od podstaw, co pokazuje, że transfer uczenia z innych zbiorów danych humorystycznych pozwala na lepsze zrozumienie złożoności i subtelności humoru Jednakże, dla zbioru Jester efekty transferu uczenia były mniej wyraźne. Specyfika tego zbioru, charakteryzująca się bardziej jednolitymi treściami humorystycznymi, mogła ograniczać korzyści płynące z transferu uczenia. Mimo to, transfer uczenia przyniósł pozytywne rezultaty, co sugeruje, że może on mieć korzystny wpływ, nawet w kontekście mniej zróżnicowanych danych.

7.7.4 Wnioskowanie z wykorzystaniem modeli generatywnych

Modele generatywne, takie jak ChatGPT 3.5, stanowią istotny element współczesnych badań nad przetwarzaniem języka naturalnego (NLP). Ich zdolność do generowania tekstu na podstawie wzorców z dużych zbiorów danych sprawia, że są użyteczne w wielu zadaniach, w tym w rozpoznawaniu humoru, emocji oraz innych złożonych form komunikacji. Wnioskowanie z wykorzystaniem tych modeli obejmuje zarówno analizę danych, jak i generowanie odpowiedzi spójnych i kontekstowo adekwatnych.

Jednym z obiecujących mechanizmów, które warto zastosować w modelach generatywnych jest personalizacja poprzez przykłady *few-shot* w prompcie (*in-context learning*) - patrz zadania spersonalizowane w (Kocoń i in., 2023). Dzięki tej technice model może dostosować się do specyficznego zadania, bazując na kilku przykładach dostarczonych w promptach. W badaniach dotyczących rozpoznawania humoru, technika *in-context learning* umożliwia dostrojenie modelu do określonych wzorców humoru, bez konieczności pełnego trenowania lub dotrenowywania. Jest to efektywne podejście, szczególnie w zadaniach o ograniczonych zasobach.

Jednak, aby osiągnąć lepszą skuteczność, można zastosować bardziej zaawansowane metody douczania, takie jak *fine-tuning* z wykorzystaniem adapterów. Technika fine-tuning wymaga jednak dostępu do parametrów modelu i pozwala na głębsze dostosowanie go do specyficznych potrzeb. Niestety nie jest ona możliwa dla modeli GPT z powodu braku otwartego dostępu do parametrów modelu. Praca zespołu HumanNLP pokazuje, że użycie adapterów w fine-tuningu prowadzi do lepszych wyników w zadaniach personalizacji (Woźniak i in., 2024). Mimo większych zasobów potrzebnych do fine-tuningu, metoda ta oferuje bardziej precyzyjne wnioskowanie w porównaniu z klasycznym podejściem generatywnym.

W kontekście badań nad klasyfikacją humoru w zakresie modeli generatywnych, warto rozważyć zastosowanie personalizacji zamiast standardowej generalizacji. Wprowadzenie personalizacji w takich scenariuszach badawczych, jak rozpoznawanie humoru, może znacząco poprawić skuteczność modeli, ponieważ pozwala lepiej uchwycić indywidualne różnice w odbiorze treści humorystycznych. Modele generatywne, mimo swojej wszechstronności, mogą mieć ograniczoną skuteczność w uogólnionych zadaniach, co podkreślają wyniki eksperymentalne. Dlatego zastosowanie technik personalizacyjnych, takich jak *few-shot learning* lub *fine-tuning* z adapterami, jest kluczowe dla uzyskania wyższej precyzji w kontekstowych zadaniach rozumienia treści humorystycznych.

7.7.5 Podsumowanie dyskusji

Wyniki badań wskazują na dominację architektur GRU oraz Wielogłowicowej atencji w zadaniach związanych z rozpoznawaniem śmieszności. Transfer uczenia również wykazuje znaczący wpływ na poprawę jakości predykcji, chociaż efekty te mogą się różnić w zależności od charakterystyki danych. Wyniki te podkreślają znaczenie wyboru odpowiednich modeli oraz technik transferu uczenia w kontekście przetwarzania języka naturalnego i rozpoznawania humoru. Przeprowadzone badania nad zadaniem rozpoznawania śmieszności potwierdziły, że spersonalizowane architektury sieci neuronowych, zwłaszcza te zaprezentowane w niniejszej pracy, wykazują znacznie wyższą skuteczność w porównaniu do innych metod spersonalizowanych oraz podejścia zgeneralizowanego, co potwierdza hipotezę H2.. Modele takie jak Attention oraz zaawansowane architektury typu transformer, jak MPNet i DeBERTa, osiągnęły najwyższe wyniki we wszystkich scenariuszach i zbiorach danych. Szczególnie dobrze radziły sobie z rozpoznawaniem bardziej złożonych form humoru, takich jak ironia, sarkazm czy gry słowne, co było trudne do uchwycenia dla prostszych modeli, takich jak CBOW czy Skipgram.

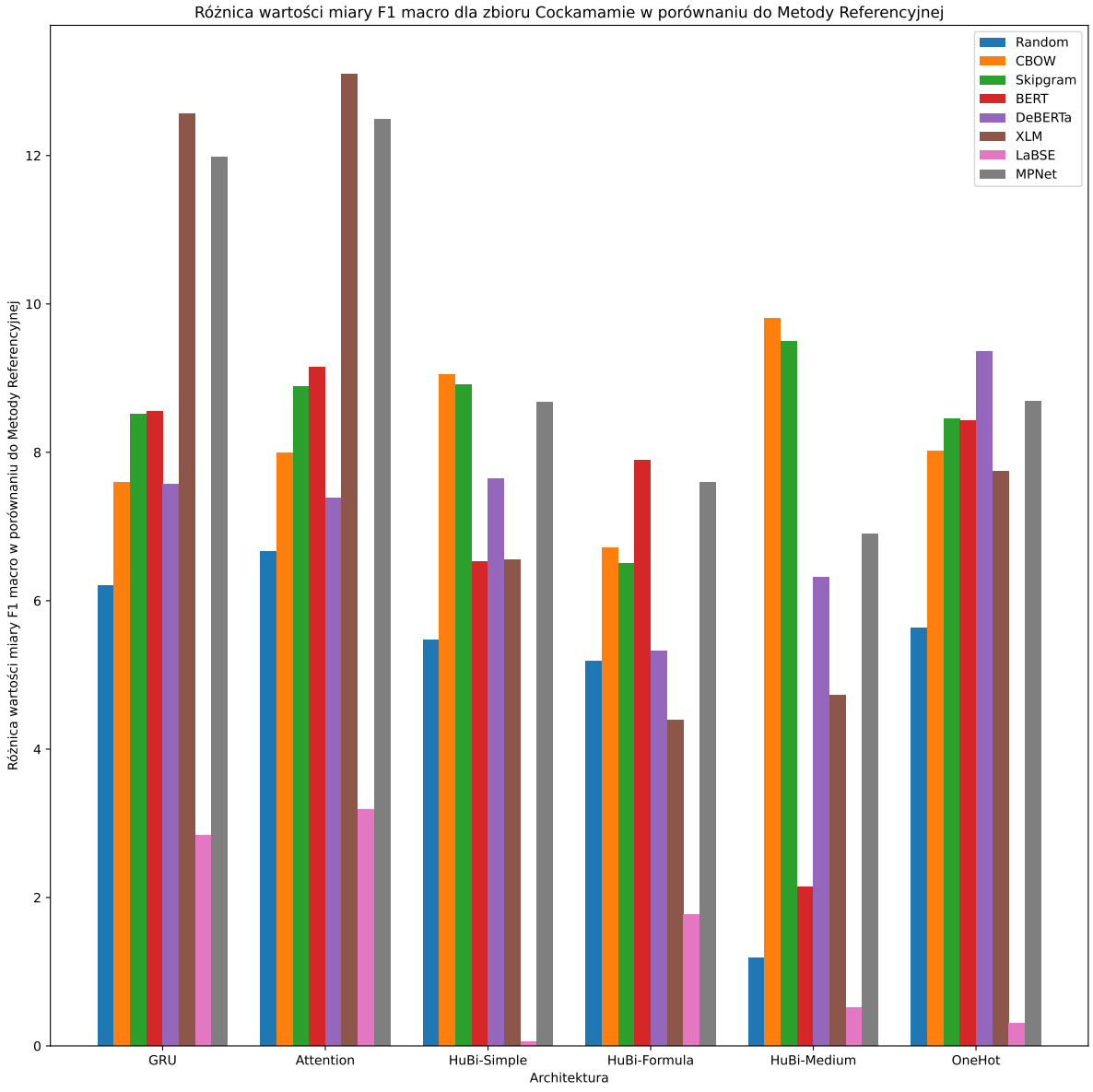
Wyniki dla różnych zbiorów danych – Cockamamie, Humicroedit, Humor, Doccano1, Doccano2, oraz Jester – jednoznacznie pokazują, że personalizowane modele, zwłaszcza te wykorzystujące mechanizmy atencji i głęboką analizę kontekstową, przewyższają tradycyjne podejścia zgeneralizowane. F1 macro dla zaawansowanych modeli często przekraczało 65-70%, podczas gdy prostsze modele językowe, jak CBOW i Skipgram, uzyskiwały wyniki o kilkanaście procent niższe.

W szczególności, zaawansowane architektury neuronowe, takie jak MPNet, osiągnęły najlepsze wyniki, przewyższając inne metody personalizowane oraz zgeneralizowane. Wyniki te potwierdzają, że zastosowanie spersonalizowanych podejść, w których modele są dostosowane do specyfiki humoru i kontekstu odbiorcy, pozwala na lepsze rozpoznawanie niuansów humorystycznych. Z kolei modele zgeneralizo-

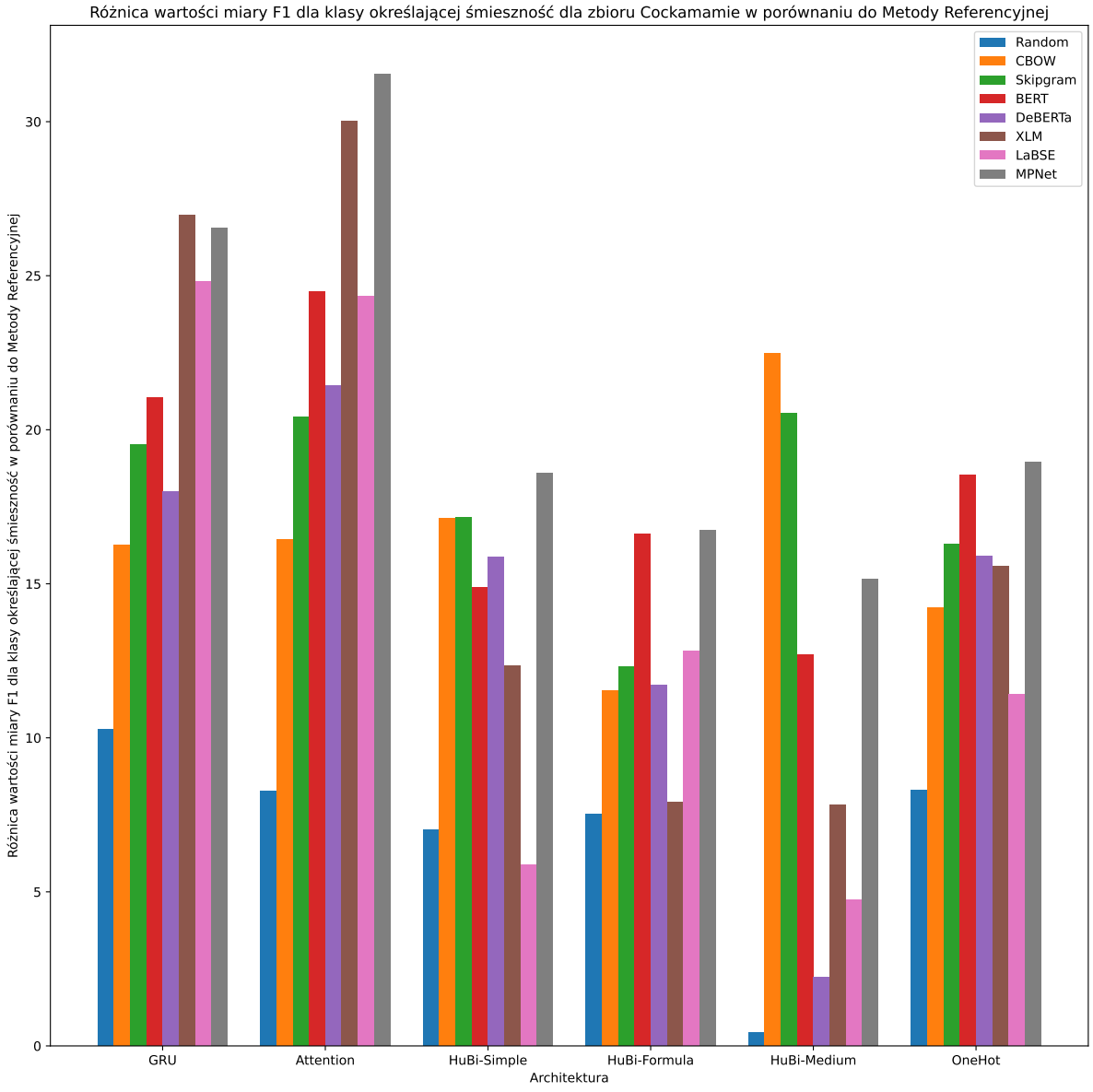
wane, mimo że oferują przyzwoitą skuteczność, nie są w stanie dorównać metodom personalizowanym, które lepiej radzą sobie z analizą wielowarstwowych i subiektywnych form humoru.

Ostatecznie, praca ta pokazuje, że personalizacja w analizie humoru jest kluczowym czynnikiem poprawiającym jakość predykcji śmieszności i otwiera nowe możliwości dla dalszych badań nad bardziej dopasowanymi do odbiorcy modelami humorystycznymi.

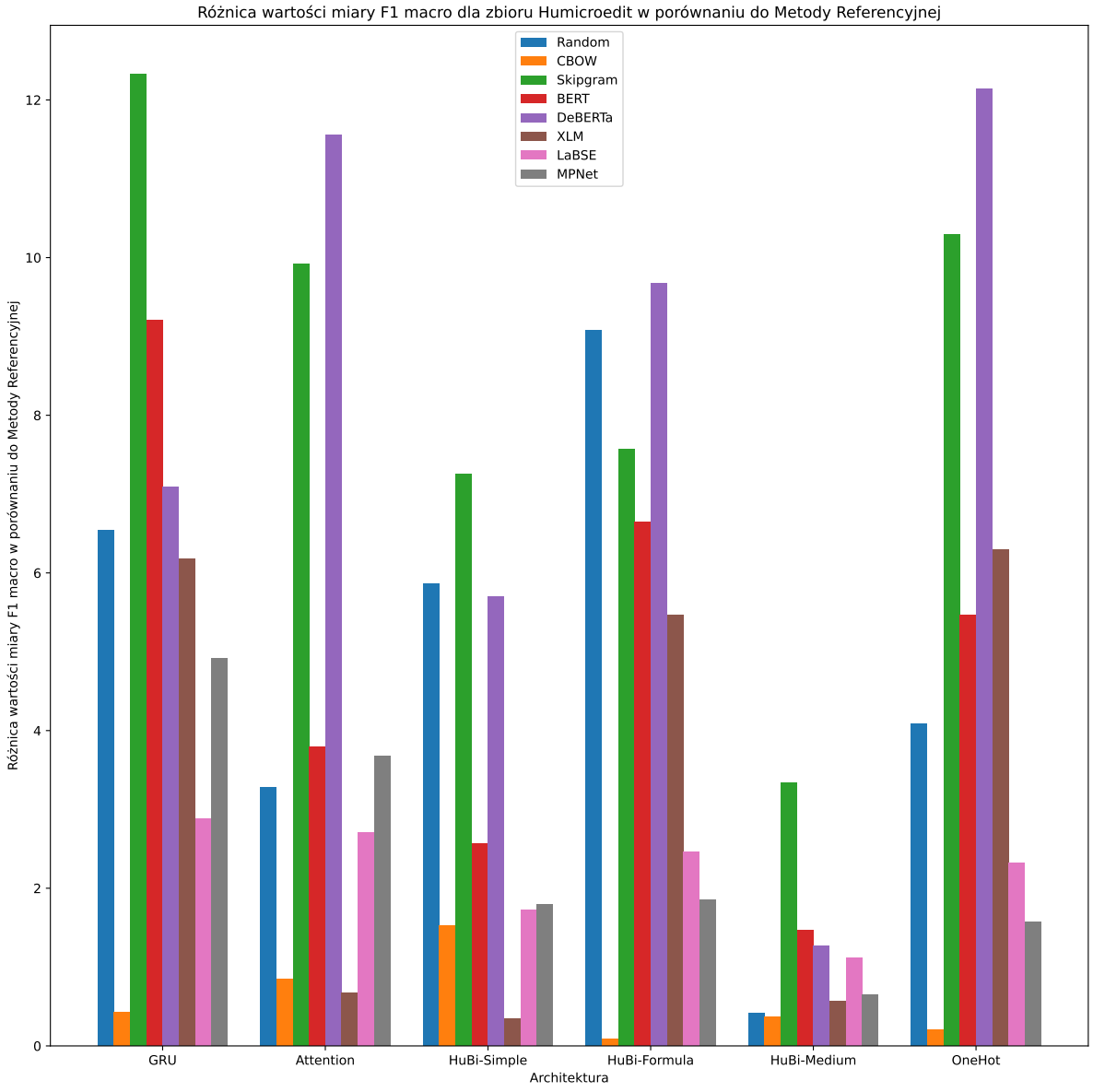
W kontekście wnioskowania z wykorzystaniem modeli generatywnych, takich jak ChatGPT 3.5, analiza wyników pokazuje, że choć modele te mają potencjał w zadaniach rozumienia humoru, ich skuteczność w bardziej złożonych przypadkach jest ograniczona w porównaniu do modeli dedykowanych (SOTA). Wprowadzenie personalizacji, zarówno poprzez in-context learning, jak i fine-tuning z adapterami, może znacząco poprawić wyniki tych modeli, lepiej dostosowując je do specyficznych zadań. W badaniach nad klasyfikacją humoru, gdzie subiektywne różnice odgrywają kluczową rolę, personalizacja okazuje się bardziej efektywna niż tradycyjna generalizacja, co stanowi ważny kierunek dalszych prac nad rozwijaniem modeli generatywnych w tym obszarze.



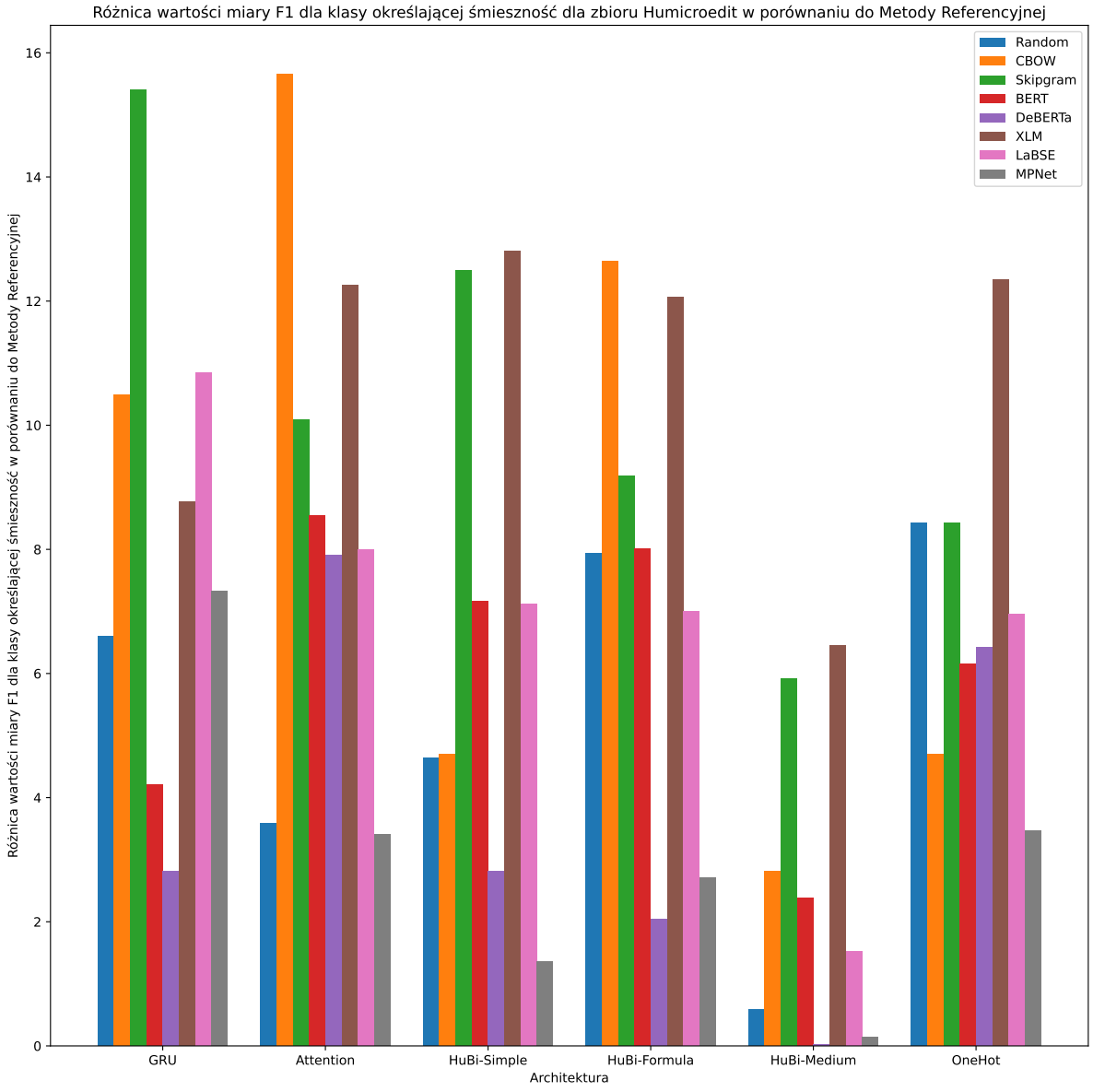
Rysunek 22: Różnica wartości miary F1 macro (p.p.) w porównaniu z Metodą Referencyjną w zależności od wykorzystanego modelu językowego dla zbioru Cockamamie. (Źródło: Opracowanie własne).



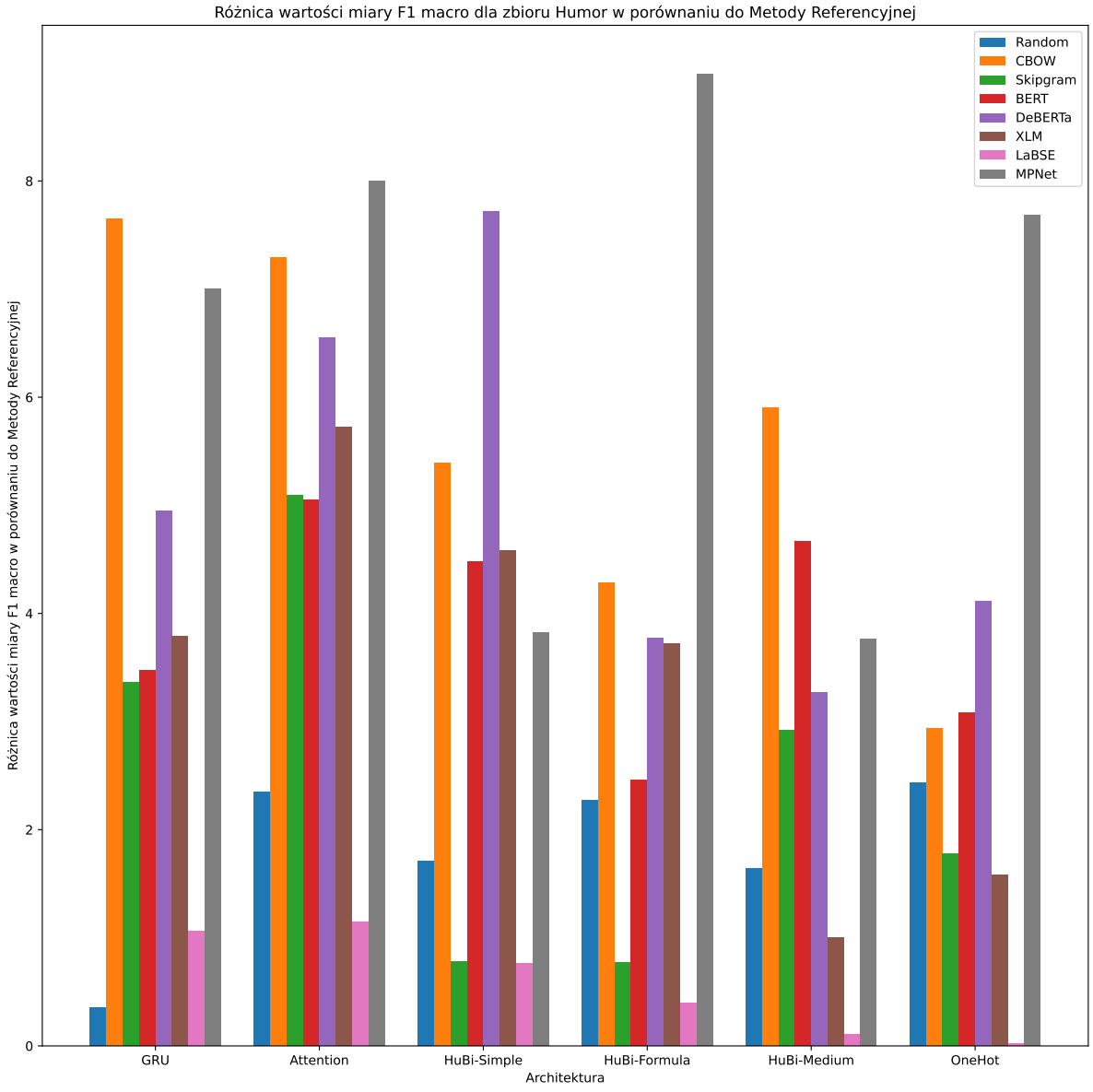
Rysunek 23: Różnica wartości miary F1 (p.p.) dla klasy oznaczającej śmieszność w porównaniu z Metodą Referencyjną w zależności od wykorzystanego modelu językowego dla zbioru Cockamamie. (Źródło: Opracowanie własne).



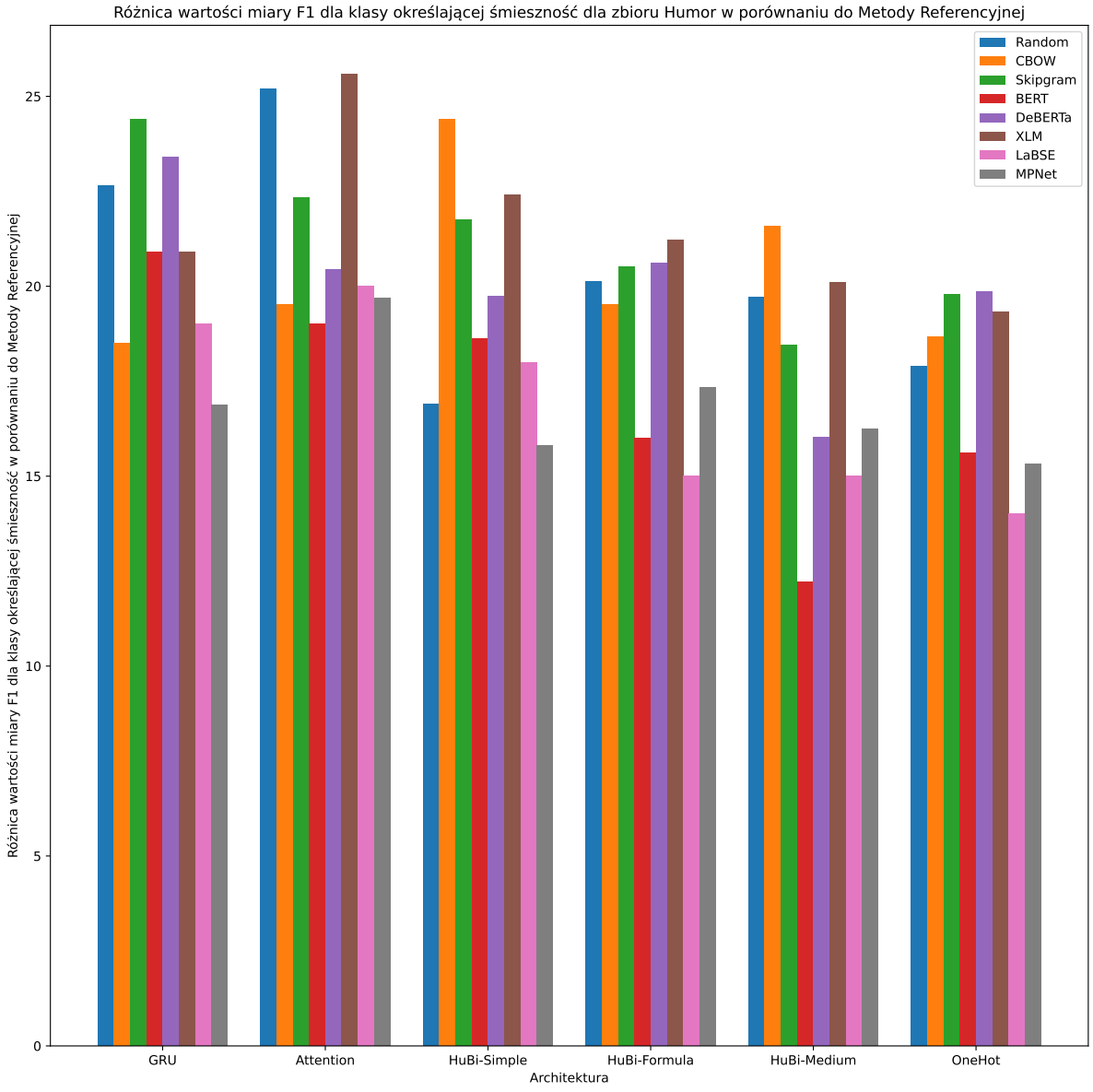
Rysunek 24: Różnica wartości miary F1 macro (p.p.) w porównaniu z Metodą Referencyjną w zależności od wykorzystanego modelu językowego dla zbioru Humicroedit. (Źródło: Opracowanie własne).



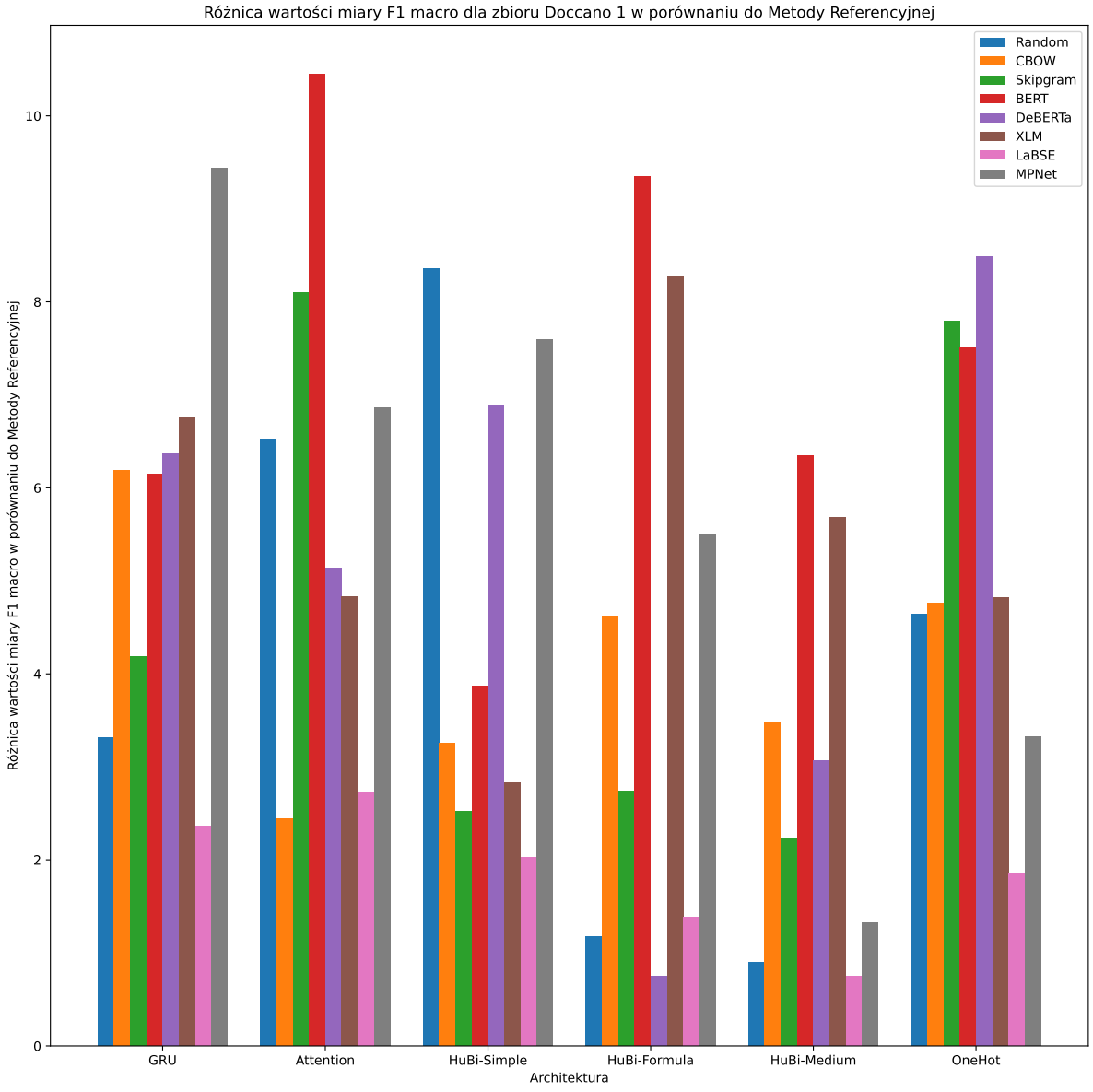
Rysunek 25: Różnica wartości miary F1 (p.p.) dla klasy oznaczającej śmieszność w porównaniu z Metodą Referencyjną w zależności od wykorzystanego modelu językowego dla zbioru Humicroedit. (Źródło: Opracowanie własne).



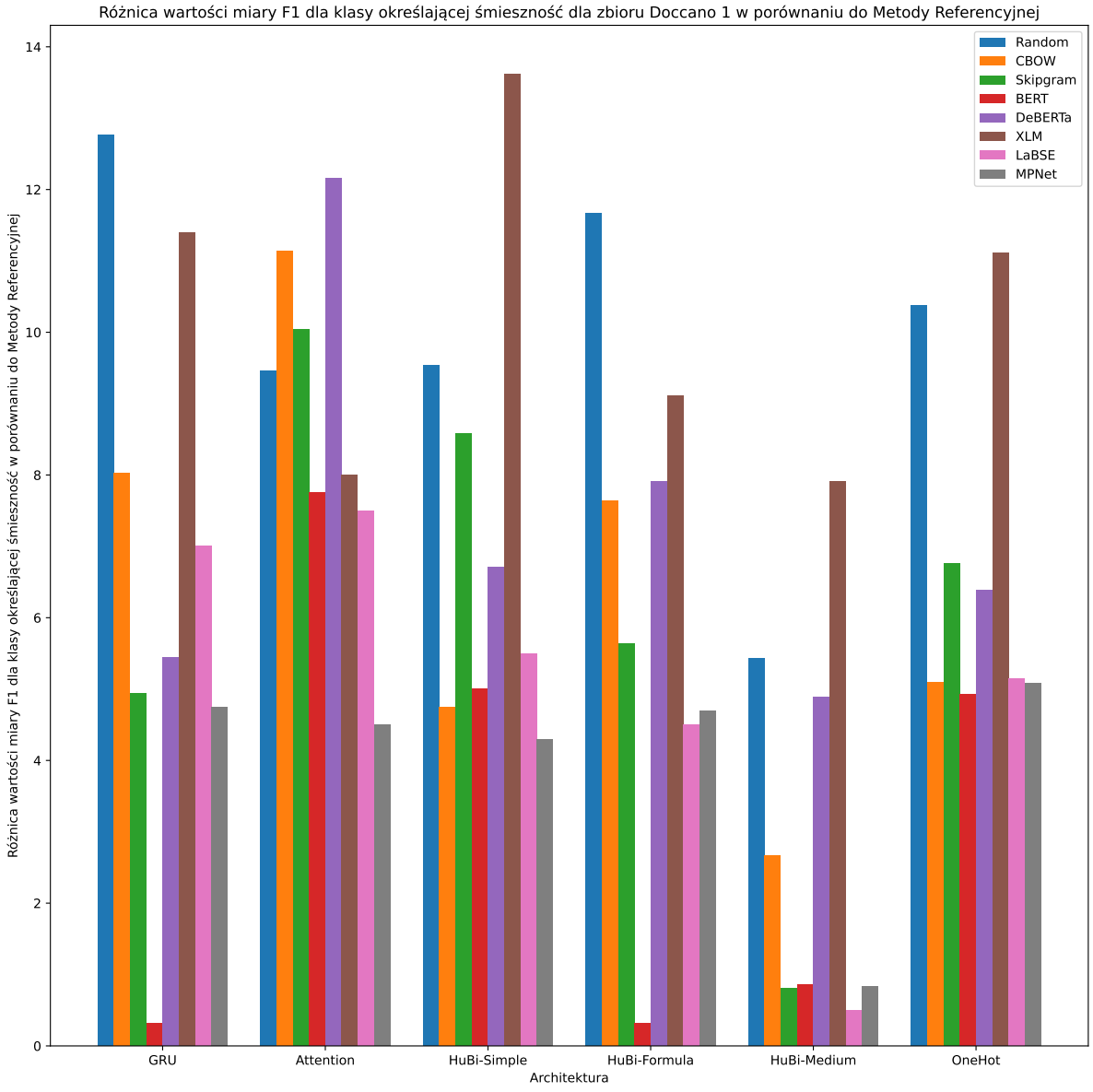
Rysunek 26: Różnica wartości miary F1 macro (p.p.) w porównaniu z Metodą Referencyjną w zależności od wykorzystanego modelu językowego dla zbioru Humor. (Źródło: Opracowanie własne).



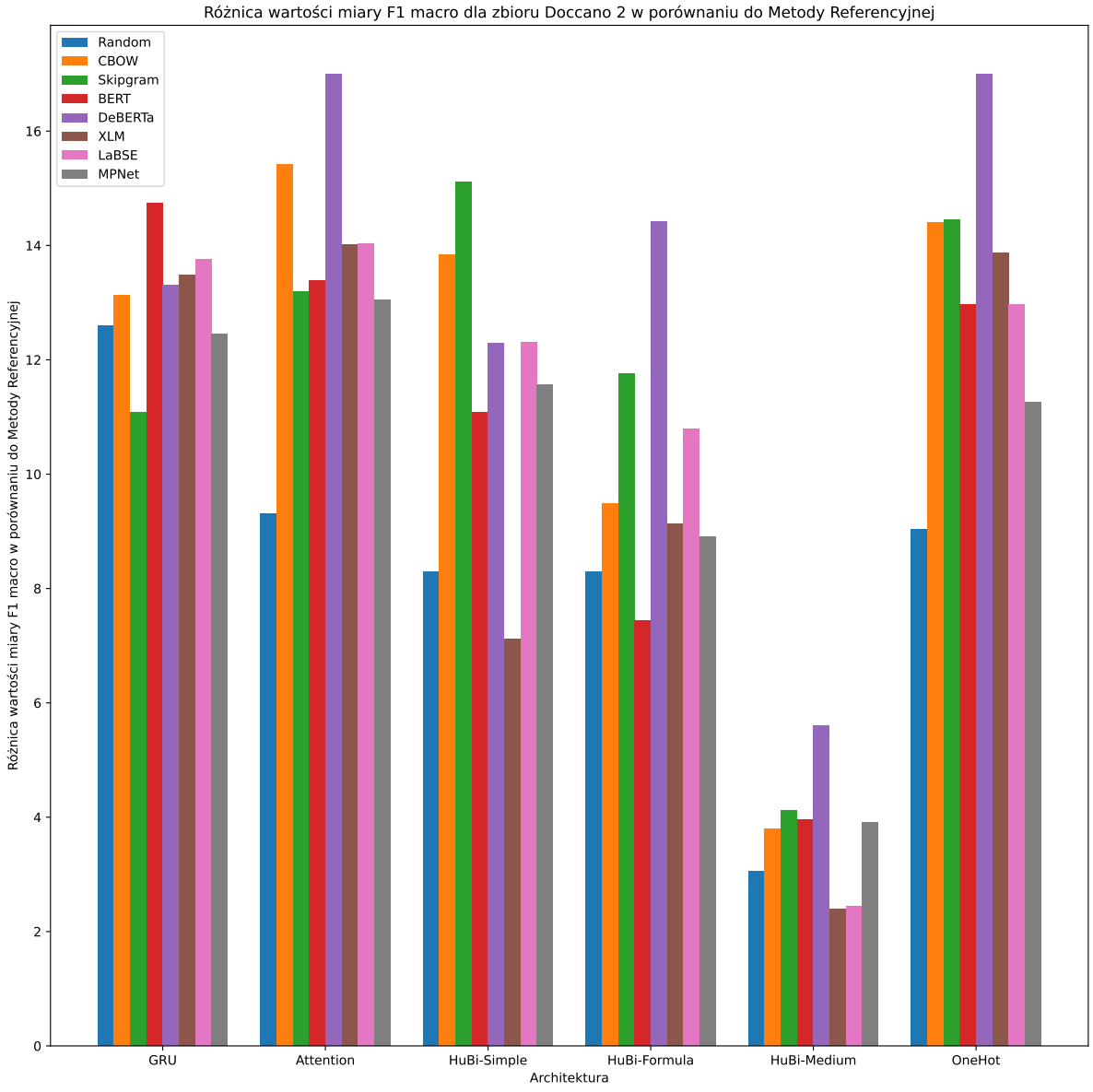
Rysunek 27: Różnica wartości miary F1 (p.p.) dla klasy oznaczającej śmieszność w porównaniu z Metodą Referencyjną w zależności od wykorzystanego modelu językowego dla zbioru Humor. (Źródło: Opracowanie własne).



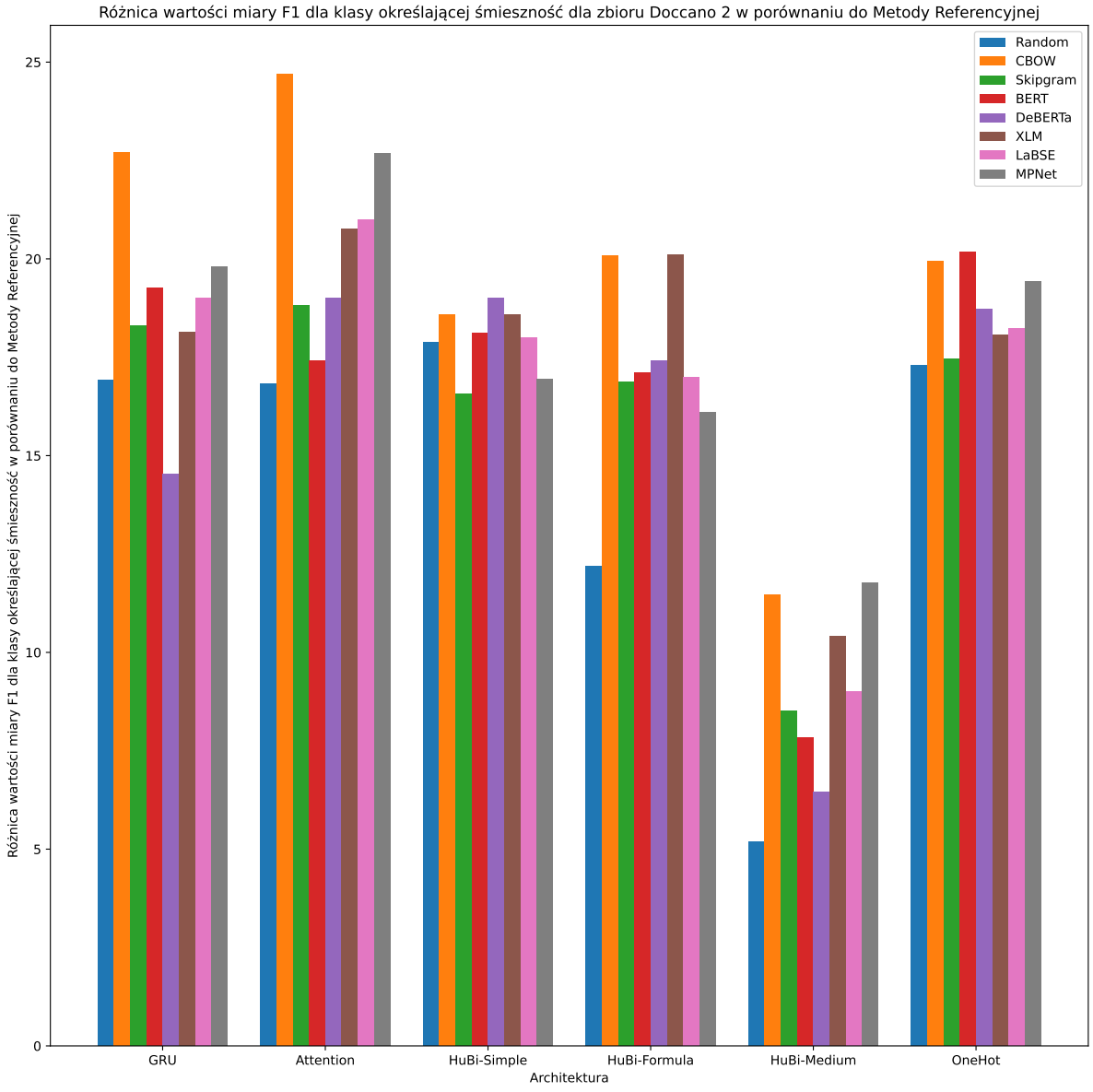
Rysunek 28: Różnica wartości miary F1 macro (p.p.) w porównaniu z Metodą Referencyjną w zależności od wykorzystanego modelu językowego dla zbioru Doccano 1. (Źródło: Opracowanie własne).



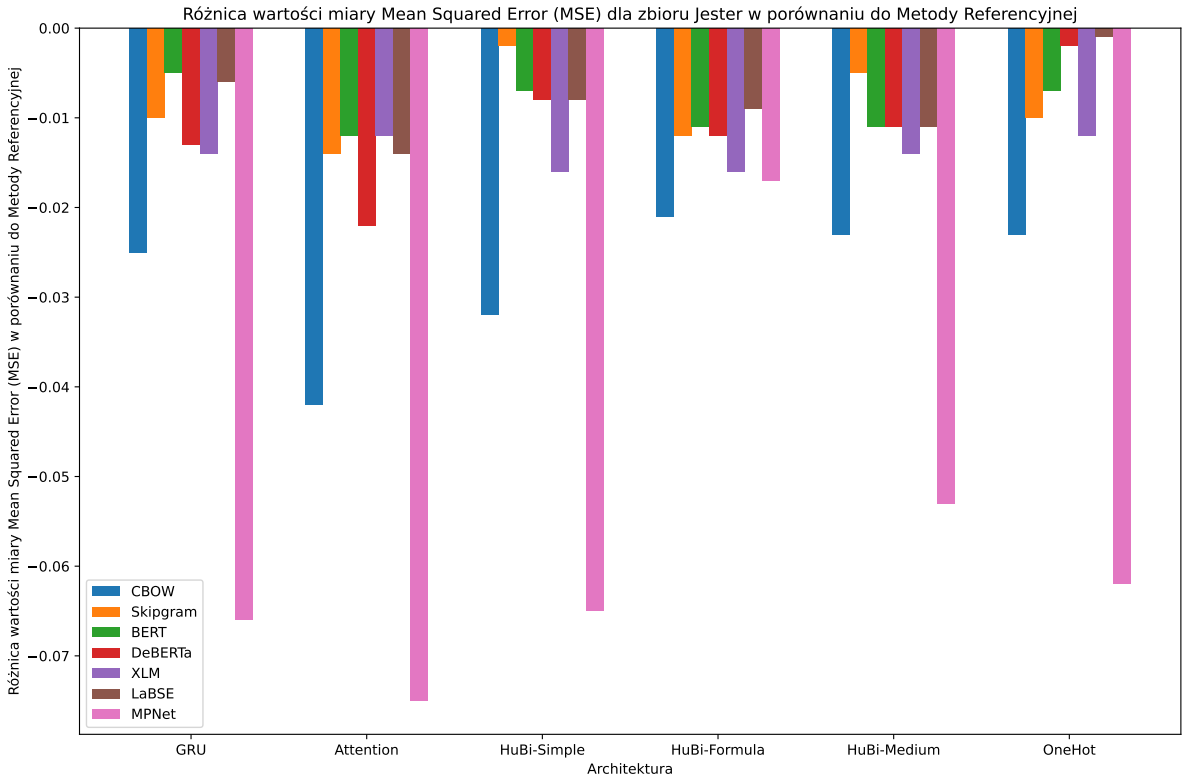
Rysunek 29: Różnica wartości miary F1 (p.p.) dla klasy oznaczającej śmieszność w porównaniu z Metodą Referencyjną w zależności od wykorzystanego modelu językowego dla zbioru Doccano 1. (Źródło: Opracowanie własne).



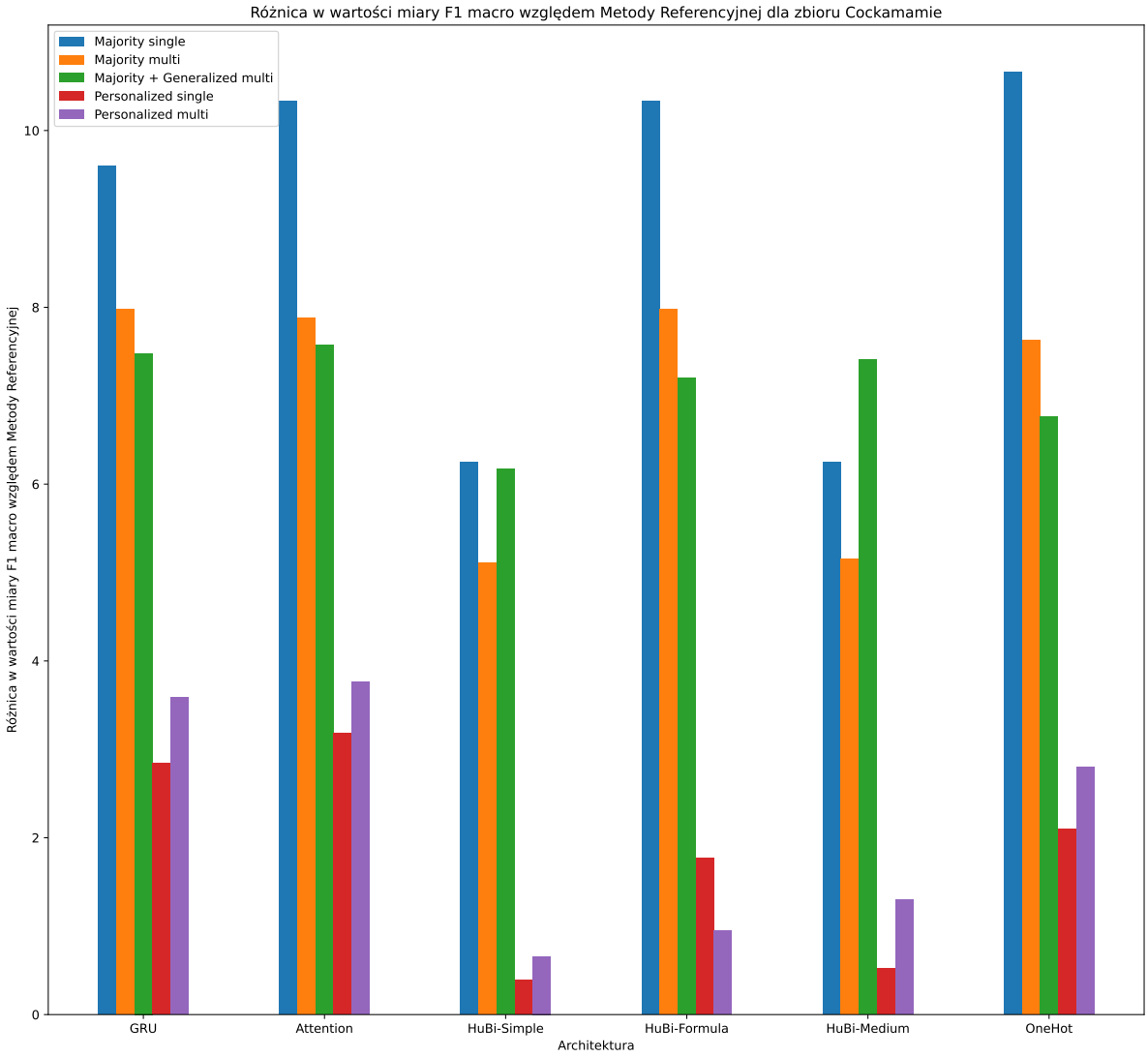
Rysunek 30: Różnica wartości miary F1 macro (p.p.) w porównaniu z Metodą Referencyjną w zależności od wykorzystanego modelu językowego dla zbioru Doccano 2. (Źródło: Opracowanie własne).



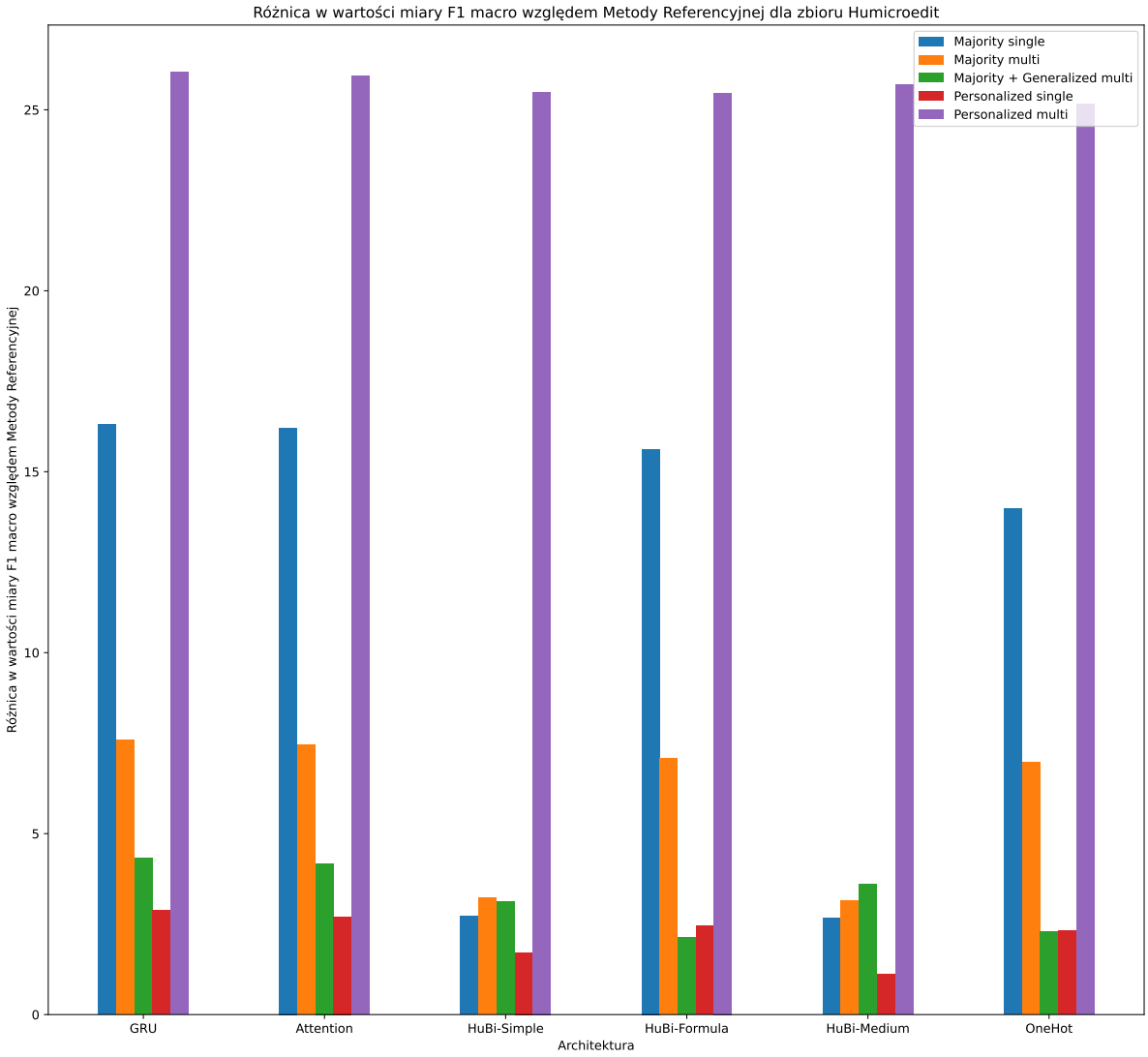
Rysunek 31: Różnica wartości miary F1 (p.p.) dla klasy oznaczającej śmieszność w porównaniu z Metodą Referencyjną w zależności od wykorzystanego modelu językowego dla zbioru Doccano 2. (Źródło: Opracowanie własne).



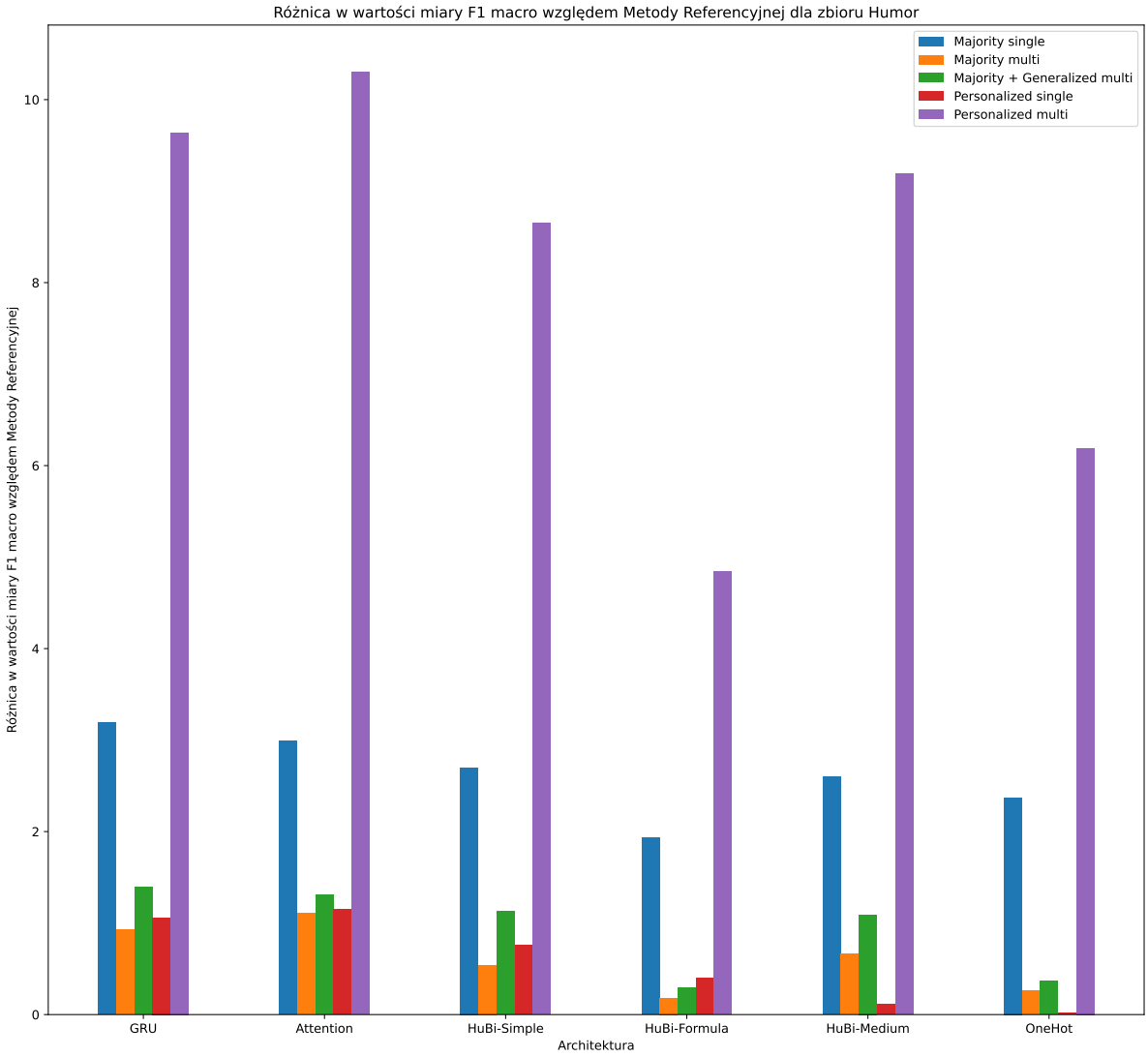
Rysunek 32: Różnica wartości miary Mean Squared Error (MSE) (p.p.) w porównaniu z Metodą Referencyjną w zależności od wykorzystanego modelu językowego dla zbioru Jester. (Źródło: Opracowanie własne).



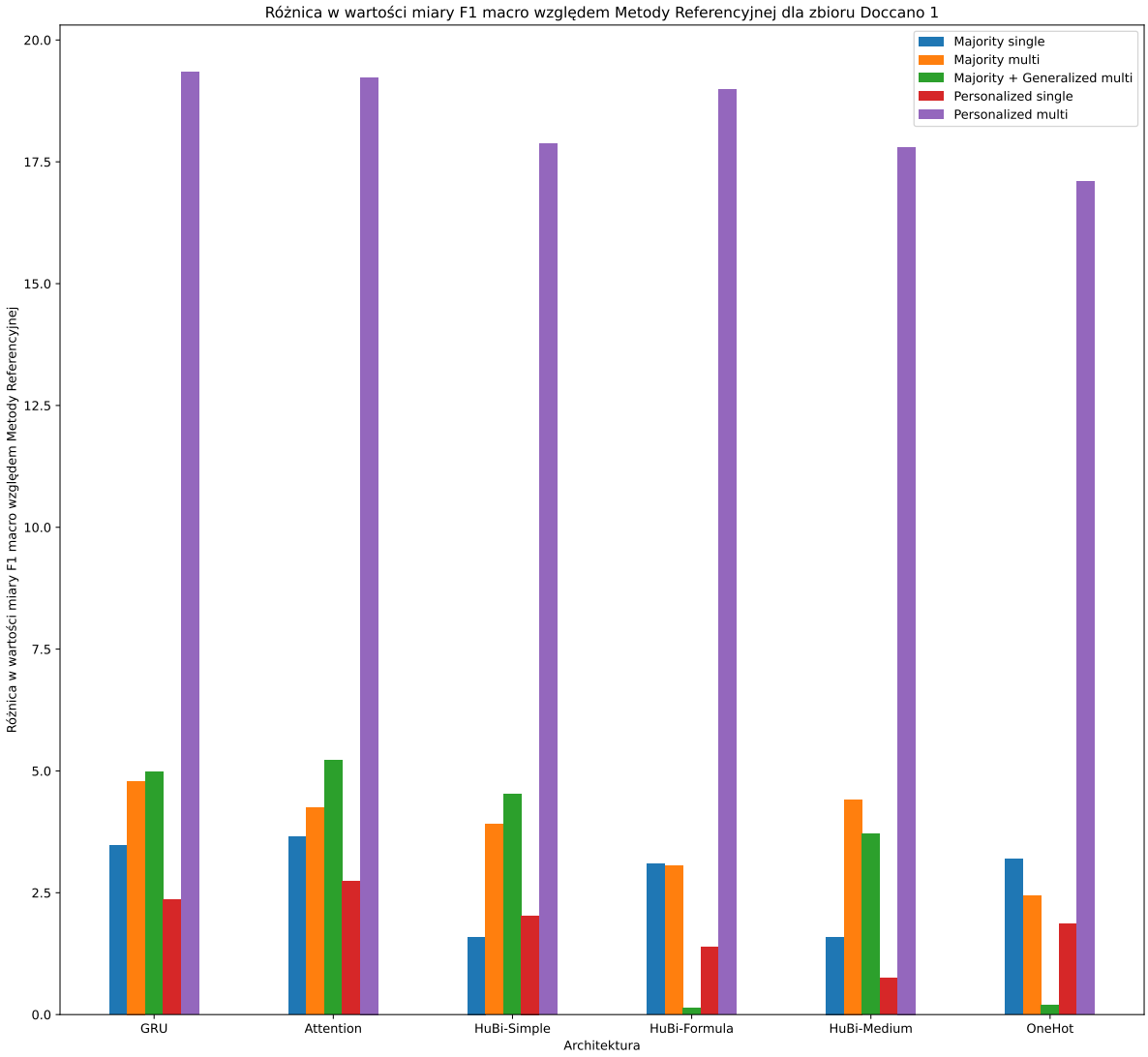
Rysunek 33: Różnice w wartościach miary F1 macro względem Metody Referencyjnej w zależności od zastosowanej strategii transferu uczenia dla zbioru Cockamamie. (Źródło: Opracowanie własne).



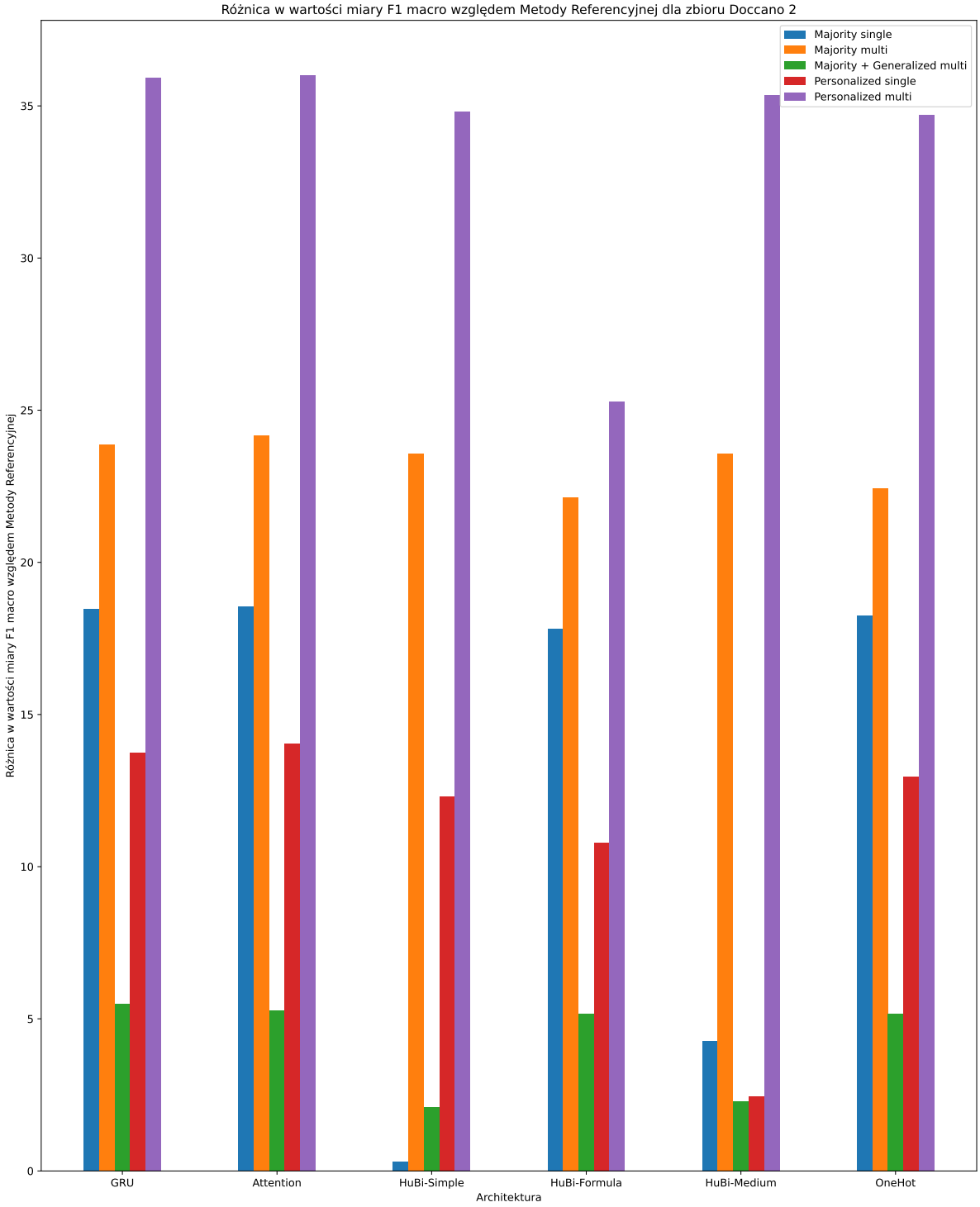
Rysunek 34: Różnice w wartościach miary F1 macro względem Metody Referencyjnej w zależności od zastosowanej strategii transferu uczenia dla zbioru Humicroedit. (Źródło: Opracowanie własne).



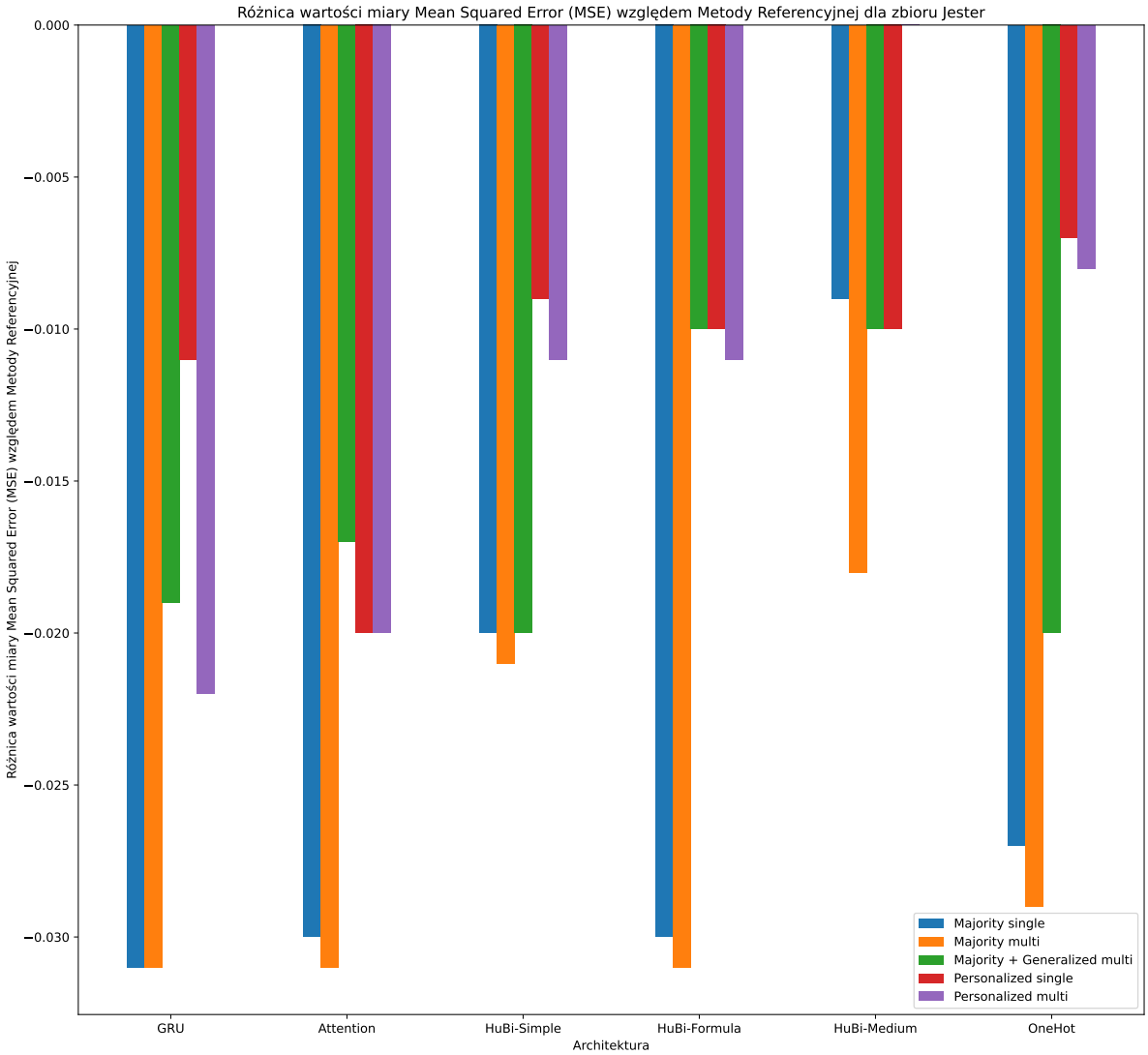
Rysunek 35: Różnice w wartościach miary F1 macro względem Metody Referencyjnej w zależności od zastosowanej strategii transferu uczenia dla zbioru Humor. (Źródło: Opracowanie własne).



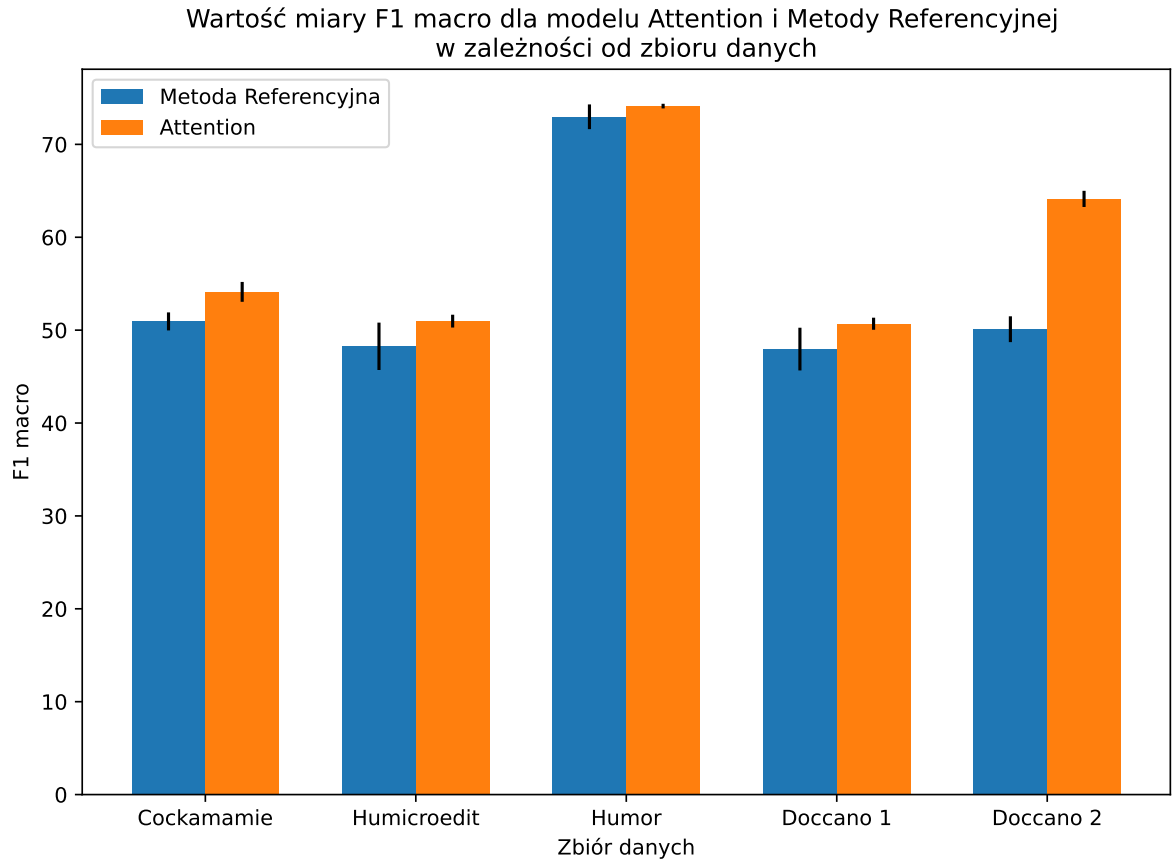
Rysunek 36: Różnice w wartościach miary F1 macro względem Metody Referencyjnej w zależności od zastosowanej strategii transferu uczenia dla zbioru Doccano 1. (Źródło: Opracowanie własne).



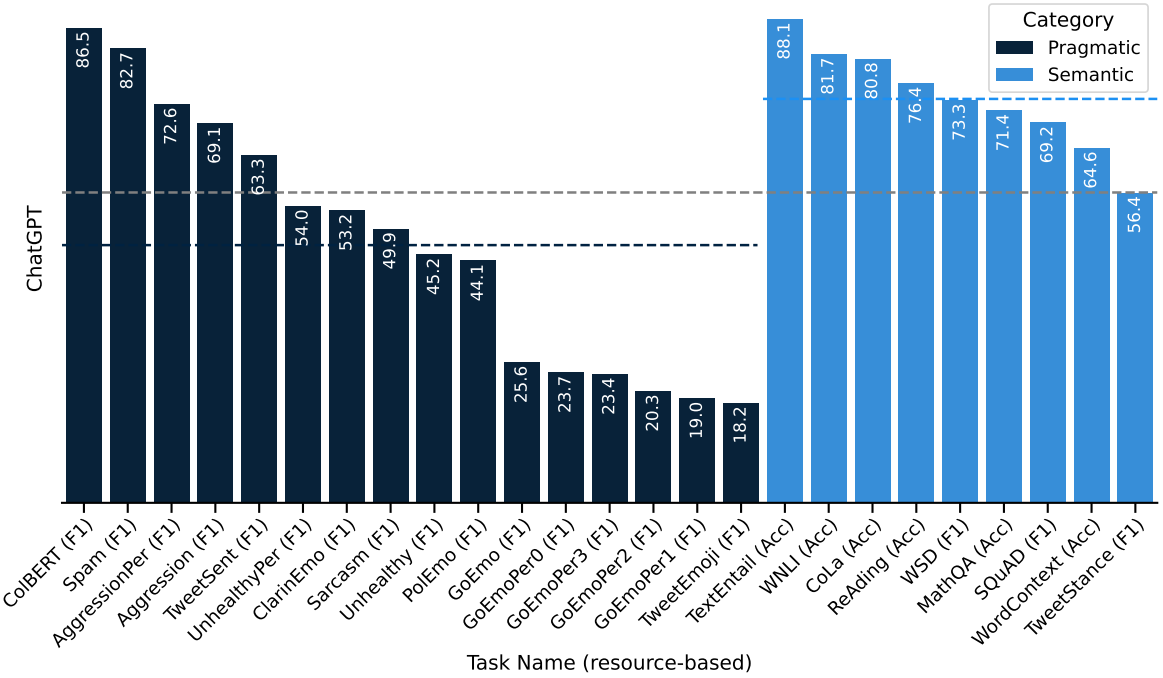
Rysunek 37: Różnice w wartościach miary F1 macro względem Metody Referencyjnej w zależności od zastosowanej strategii transferu uczenia dla zbioru Doccano 2. (Źródło: Opracowanie własne).



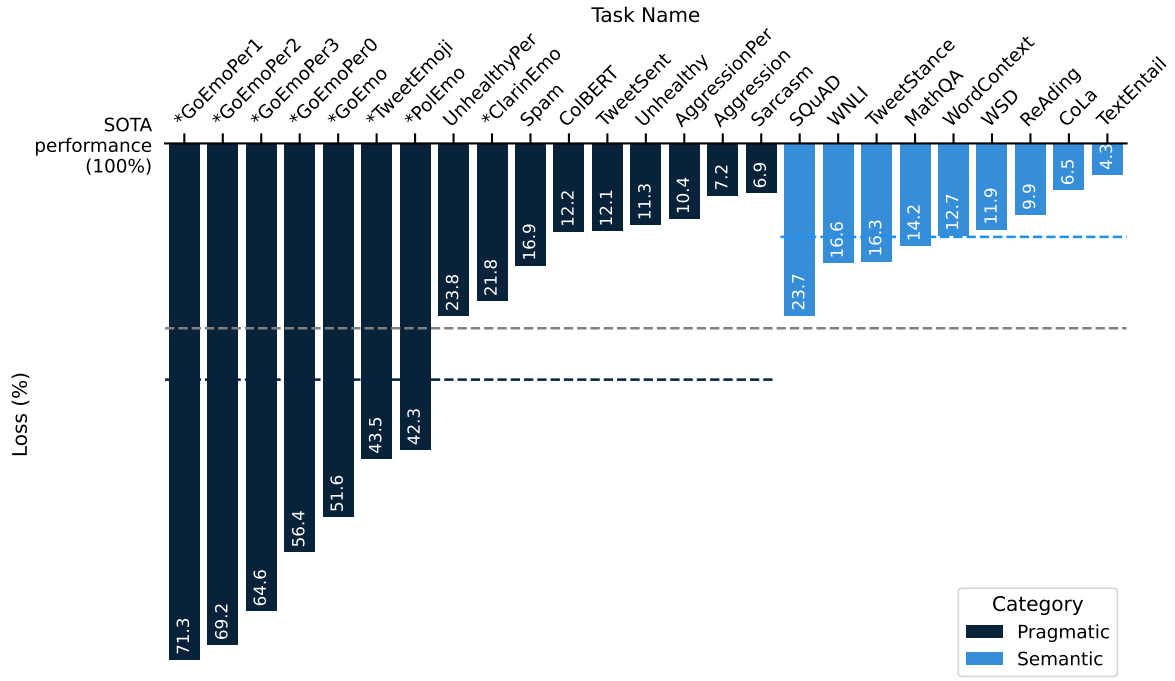
Rysunek 38: Różnice w wartościach miary Mean Squared Error (MSE) względem Metody Referencyjnej w zależności od zastosowanej strategii transferu uczenia dla zbioru Jester. (Źródło: Opracowanie własne).



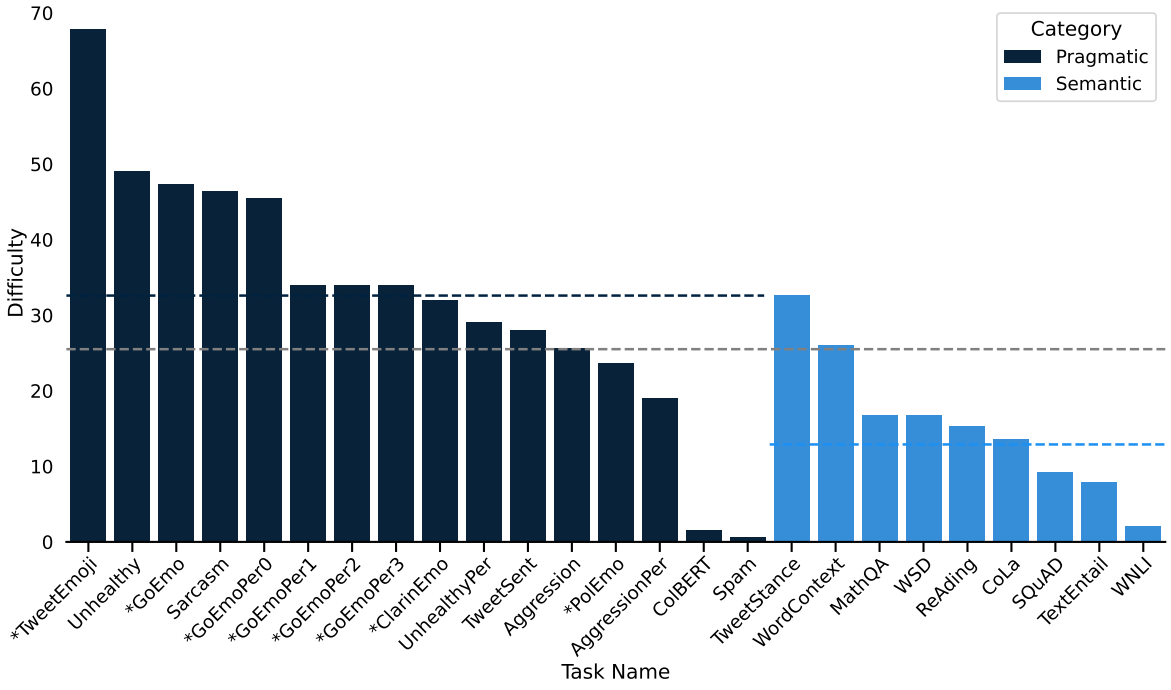
Rysunek 39: Wartość miary F1 macro dla modelu Wielogłowicowej atencji oraz Modelu Referencyjnego w zależności od zbioru danych. (Źródło: Opracowanie własne).



Rysunek 40: Wydajność modelu ChatGPT (%) dla wszystkich rozważanych zadań, nazwanych zgodnie z ich zasobami (zbiorami danych), w tym również dla zadań związanych z humorem ColBERT i Sarcasm. Przerywane linie oznaczają średnią wydajność tylko dla zadań semantycznych, wszystkich zadań oraz tylko dla zadań pragmatycznych. (Źródło: (Kocoń i in., 2023)).



Rysunek 41: Strata wydajności (Loss) modelu ChatGPT (%) dla wszystkich rozważanych zadań, nazwanych zgodnie z ich zasobami (zbiorami danych), w tym również dla zadań związanych ze śmiesznością: ColBERT i Sarcasm, uporządkowanych malejąco według wartości straty. Zadania poprzedzone gwiazdką dotyczą emocji. Górna oś X odpowiada wydajności najlepszego modelu (SOTA), traktowanego jako 100% możliwości. Przerywane linie oznaczają średnie wartości straty dla tylko zadań pragmatycznych, tylko zadań semantycznych oraz wszystkich zadań. (Źródło: (Kocoń i in., 2023)).



Rysunek 42: Trudność zadania (100% - wydajność SOTA) uporządkowana malejąco. Zadania poprzedzone gwiazdką dotyczą emocji. Przerywane linie oznaczają średni poziom trudności tylko dla zadań pragmatycznych, wszystkich zadań oraz tylko zadań semantycznych. (Źródło: (Kocoń i in., 2023)).

8 PODSUMOWANIE

Niniejsza rozprawa doktorska dotyczyła zagadnienia personalizacji w zadaniu rozpoznawania treści humorystycznych w kontekście przetwarzania języka naturalnego (NLP). Badania przeprowadzone w tej pracy wypełniają istotną lukę w dziedzinie, gdzie większość dotychczasowych podejść skupiała się na uogólnionych modelach, pomijając subiektywność poczucia humoru. W ramach pracy postawiono kilka hipotez, które miały na celu zbadanie wpływu personalizacji na jakość rozpoznawania treści humorystycznych oraz zgłębienie różnych aspektów humoru w kontekście NLP.

Przeprowadzone badania wykazały, że personalizacja znacząco poprawia skuteczność predykcji w porównaniu z podejściem uogólnionym. Modele uwzględniające indywidualne preferencje użytkowników, zwłaszcza te zaproponowane w pracy, takie jak GRU oraz mechanizmy wielogłowicowej atencji, pozwoliły na bardziej precyzyjne dopasowanie ocen humorystycznych do unikalnych doświadczeń i preferencji odbiorców.

GRU było lepsze niż inne architektury od innych spersonalizowanych od x do y a od zgeneralizowanej architektury od x do y w zalezności od scenariusza eksperymentalnego.

Dodatkowo, w pracy opracowano szczegółową kategoryzację humoru, obejmującą różnorodne formy, takie jak ironia, sarkazm, gry słowne i absurd, co umożliwiło budowę bardziej zaawansowanych systemów rozpoznawania humoru i głębszą analizę jakości wnioskowania dla różnych rodzajów humoru.

Szczególną uwagę poświęcono zastosowaniu zaawansowanych modeli językowych, takich jak architektury typu Transformer, które okazały się skuteczniejsze w rozpoznawaniu humoru niż prostsze podejścia, takie jak modele CBOW (Continuous Bag of Words) czy Skip-gram. Badania pokazały, że bardziej zaawansowane techniki reprezentacji tekstu pozwalają na uchwycenie subtelniejszych różnic w humorze, co przyczynia się do poprawy wyników w zadaniach klasyfikacyjnych.

Ważnym elementem pracy było również badanie wpływu rozmiaru zbiorów danych oraz uczenia modeli na tzw. foldach (krokach w walidacji krzyżowej). Analiza wyników uzyskanych na różnych liczbach foldów pozwoliła na ocenę stabilności i generalizacyjności modeli w różnych warunkach eksperymentalnych. Wyniki potwierdziły, że zwiększenie liczby foldów prowadzi do poprawy jakości predykcji, co pozwala na lepsze uogólnienie wyników w przypadku nowych danych, przy czym szybkość zwiekszania jakości zależy od charakteru humoru.

W pracy zastosowano również techniki transferu uczenia, które umożliwiły łączenie różnych zbiorów trenigowych lub wykorzystanie informacji pozyskanych w trakcie trenowania na jednym zbiorze podczas wnioskowania na innym zbiorze. W efekcie transfer uczenia zwiększał efektywności modeli oraz poprawiał ich zdolności do generalizacji.

W przeprowadzonych badaniach nad modelem ChatGPT 3.5 skupiono się na analizie jego skuteczności w zadaniach związanych z klasyfikacją humoru w porównaniu do modeli dedykowanych (SOTA). Wykazano, że choć modele generatywne, takie jak ChatGPT, osiągają satysfakcjonujące wyniki w prostszych zadaniach, w przypadku bardziej złożonych kontekstów humorystycznych, np. sarkazmu, modele dedykowane zdecydowanie przeważają pod względem skuteczności. Ponadto, personalizacja poprzez *in-context learning* zastosowana w innych zadaniach oraz douczanie (*fine-tuning*) z wykorzystaniem adapterów pozwala znacząco poprawić jakość predykcji (Kocoń i in., 2023), (Woźniak i in., 2024), co sugeruje, że warto w przyszłych badaniach kłaść większy nacisk na te techniki, zamiast polegać wyłącznie na ogólnej generalizacji.

Podsumowując, praca wnosi istotny wkład w rozwój dziedziny przetwarzania języka naturalnego, pokazując, że personalizacja, wykorzystanie zaawansowanych modeli językowych oraz technik transferu uczenia są kluczowe dla precyzyjnej analizy i generowania treści humorystycznych. Opracowane metody mogą znaleźć szerokie zastosowanie nie tylko w analizie humoru, ale także w innych zadaniach NLP, gdzie subiektywne preferencje użytkowników odgrywają istotną rolę.

9 KIERUNKI DALSZYCH PRAC

Wyniki zaprezentowane w niniejszej rozprawie otwierają szereg możliwości rozwoju i dalszych badań w dziedzinie przetwarzania języka naturalnego (NLP) oraz rozpoznawania humoru. Mimo iż badania potwierdziły skuteczność zaproponowanych metod personalizacji i zaawansowanych modeli językowych, istnieje wiele obszarów, które można dalej zgłębiać i optymalizować.

Jednym z głównych kierunków dalszych badań jest rozwijanie modeli personalizacji, szczególnie w zakresie jeszcze głębszego dopasowywania treści do indywidualnych użytkowników. Przyszłe prace mogą skoncentrować się na bardziej zaawansowanych sposobach zbierania i modelowania danych o preferencjach użytkowników. Zastosowanie bardziej złożonych metod modelowania profilu użytkownika, takich jak adaptacyjne algorytmy uczenia maszynowego, które dynamicznie dostosowują się do zmian preferencji w czasie, może zwiększyć skuteczność rozpoznawania humoru i innych subiektywnych treści.

Dalsze badania mogłyby także skupić się na personalizacji wielowymiarowej, która uwzględnia nie tylko preferencje humorystyczne, ale także inne aspekty komunikacji, takie jak nastrój, kontekst kulturowy czy sytuacyjny. Wprowadzenie bardziej zaawansowanych metod analizy emocji, które uwzględniają kontekst i dynamikę interakcji użytkowników, mogłoby zwiększyć dokładność predykcji.

Kolejnym obszarem, który warto rozważyć, jest wykorzystanie modeli multimodalnych do rozpoznawania humoru. Humor nie zawsze jest ograniczony do samego tekstu – często zawiera elementy wizualne, dźwiękowe lub inne modalności, które wpływają na odbiór treści. Zastosowanie modeli, które mogą analizować dane z różnych źródeł, takich jak obrazy, wideo czy dźwięki, pozwoliłoby na bardziej kompleksową analizę treści humorystycznych.

Wprowadzenie modeli multimodalnych mogłoby być szczególnie przydatne w analizie humoru internetowego, który często łączy tekst z memami, gifami czy wideo. Rozwój takich narzędzi mógłby rozszerzyć możliwości analizy humoru w różnych kontekstach komunikacyjnych.

Choć zastosowane techniki transferu uczenia przyniosły obiecujące rezultaty, dalsze prace mogą skupić się na optymalizacji tych metod. Istnieje potencjał w rozwijaniu bardziej efektywnych strategii transferu uczenia między różnymi domenami, językami i kulturami. Badania nad technikami transferu uczenia mogłyby uwzględniać zaawansowane podejścia, takie jak uczenie wstępne (pre-training) na bardzo du-

żych, wielojęzycznych zbiorach danych oraz dostrajanie (fine-tuning) na mniejszych, specyficznych dla zadania zbiorach.

Interesującym kierunkiem mogłoby być również badanie transferu uczenia w odniesieniu do bardzo różnorodnych typów humoru. Przykładowo, modele mogłyby być trenowane na danych pochodzących z różnych kultur lub grup społecznych, a następnie wykorzystywane do predykcji humoru w zupełnie innych kontekstach. Badania w tym zakresie mogłyby również obejmować analizę możliwości przenoszenia wiedzy pomiędzy różnymi typami treści, takimi jak komedia sytuacyjna, humor abstrakcyjny, ironia, sarkazm, i inne.

Kolejnym naturalnym krokiem jest zastosowanie przedstawionych metod na większych i bardziej zróżnicowanych zbiorach danych. Chociaż w pracy wykorzystano różnorodne zestawy danych, przyszłe badania mogłyby objąć większe populacje użytkowników, co pozwoliłoby na bardziej precyzyjne modelowanie subiektywnych preferencji humorystycznych. Dodatkowo, większe zbiory danych mogłyby umożliwić bardziej zaawansowane analizy, takie jak identyfikacja trendów humorystycznych w różnych regionach geograficznych, grupach wiekowych czy społecznościach internetowych.

Zbiór bardziej zróżnicowanych danych może również umożliwić zastosowanie nowych technik walidacji i lepsze zrozumienie, jak różne grupy reagują na te same treści humorystyczne. Warto także badać możliwość automatycznego generowania nowych, spersonalizowanych danych treningowych, co zwiększyłoby skuteczność modeli w rozpoznawaniu humoru w bardziej specyficznych kontekstach.

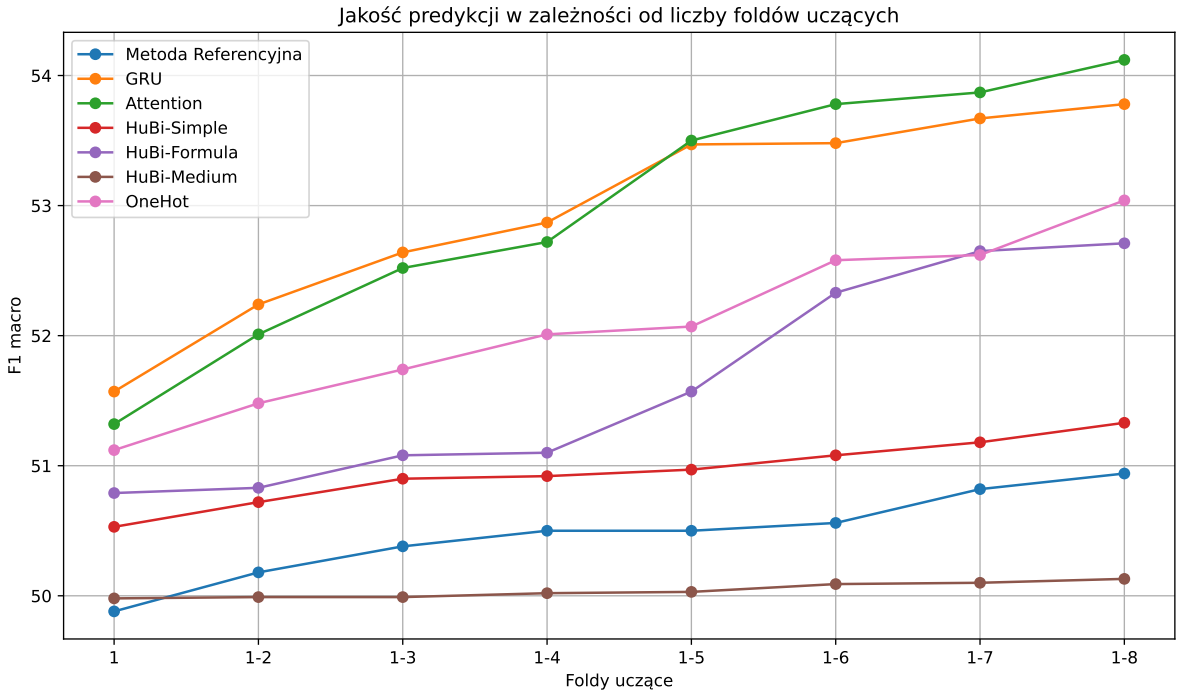
W przyszłych badaniach nad modelami generatywnymi, takimi jak ChatGPT, warto rozważyć kilka dodatkowych kierunków. Po pierwsze, integracja multimodalnych danych (np. obrazów, dźwięków) mogłaby znacząco poprawić zdolność modelu do rozpoznawania humoru, który często bazuje na kontekście wizualnym lub dźwiękowym. Po drugie, rozwój dynamicznych systemów adaptacyjnych, które na bieżąco uczą się preferencji użytkownika w czasie interakcji, mógłby zwiększyć personalizację i dokładność modelu. Innym interesującym podejściem jest eksploracja generowania kontrfaktycznego – modeli, które tworzą alternatywne scenariusze lub wersje humoru, co mogłoby poszerzyć ich zdolność do kreatywnego przetwarzania języka. Wreszcie, rozwój technik wczesnego wykrywania błędów w klasyfikacji mógłby poprawić niezawodność generatywnych modeli w zadaniach o wysokiej złożoności pragmatycznej, takich jak humor lub sarkazm.

Część IV

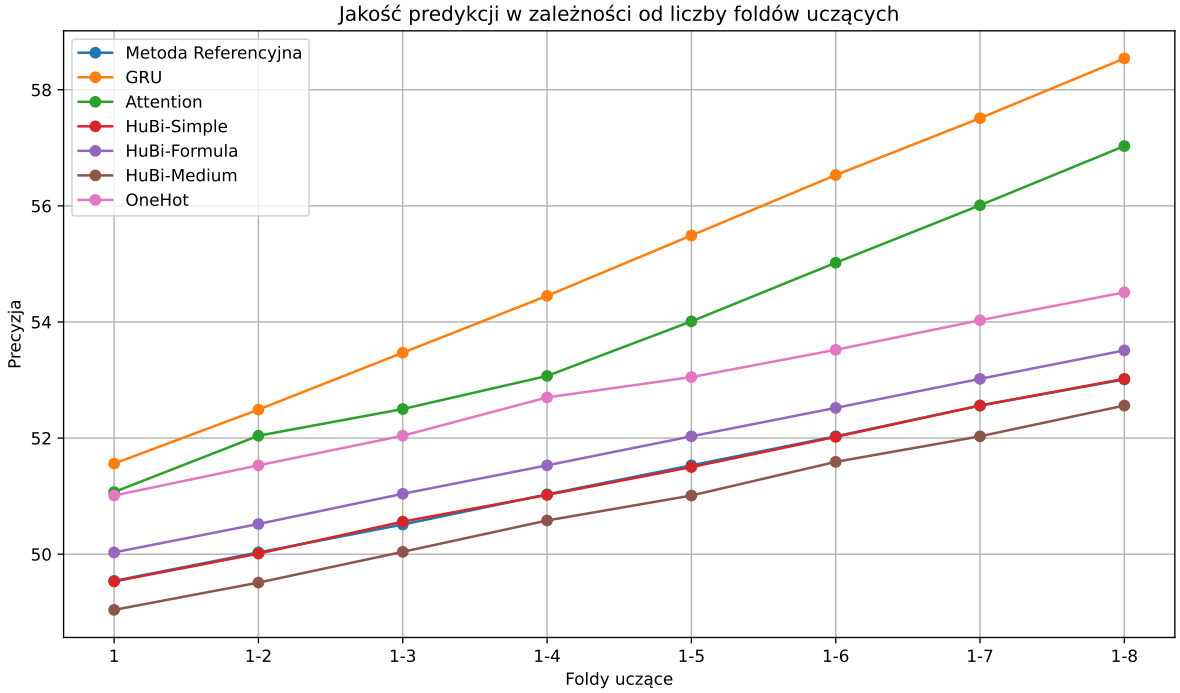
ZAŁĄCZNIKI

A SZCZEGÓŁOWE WYNIKI PRZEPROWADZONYCH EKSPERYMENTÓW

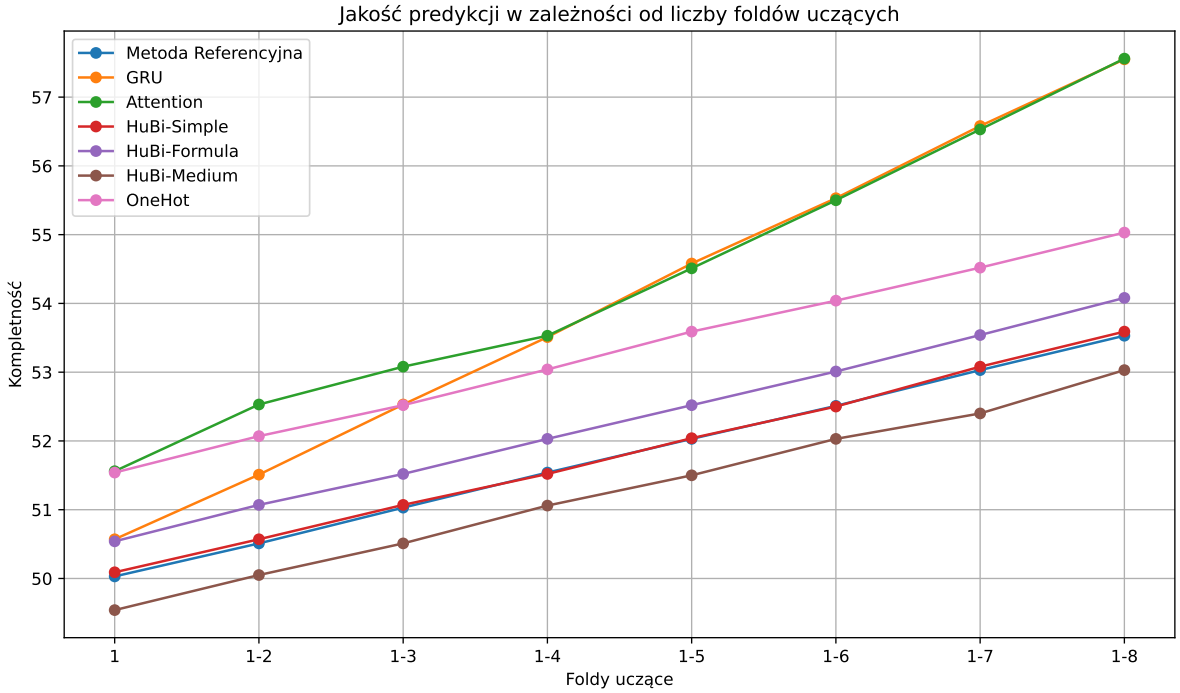
W niniejszym rozdziale przedstawiono pełne zestawienia wyników uzyskanych w trakcie badań nad predykcją śmieszności. Wyniki te, zaprezentowane w formie wykresów, szczegółowo obrazują efektywność różnych modeli językowych oraz technik, takich jak transfer uczenia, w różnych zadaniach i zestawach danych. W załączniku znajdują się wykresy ilustrujące wyniki poszczególnych eksperymentów, które umożliwiają porównanie wydajności modeli w zależności od zastosowanego podejścia, rozmiaru zbioru uczącego oraz innych kluczowych parametrów.



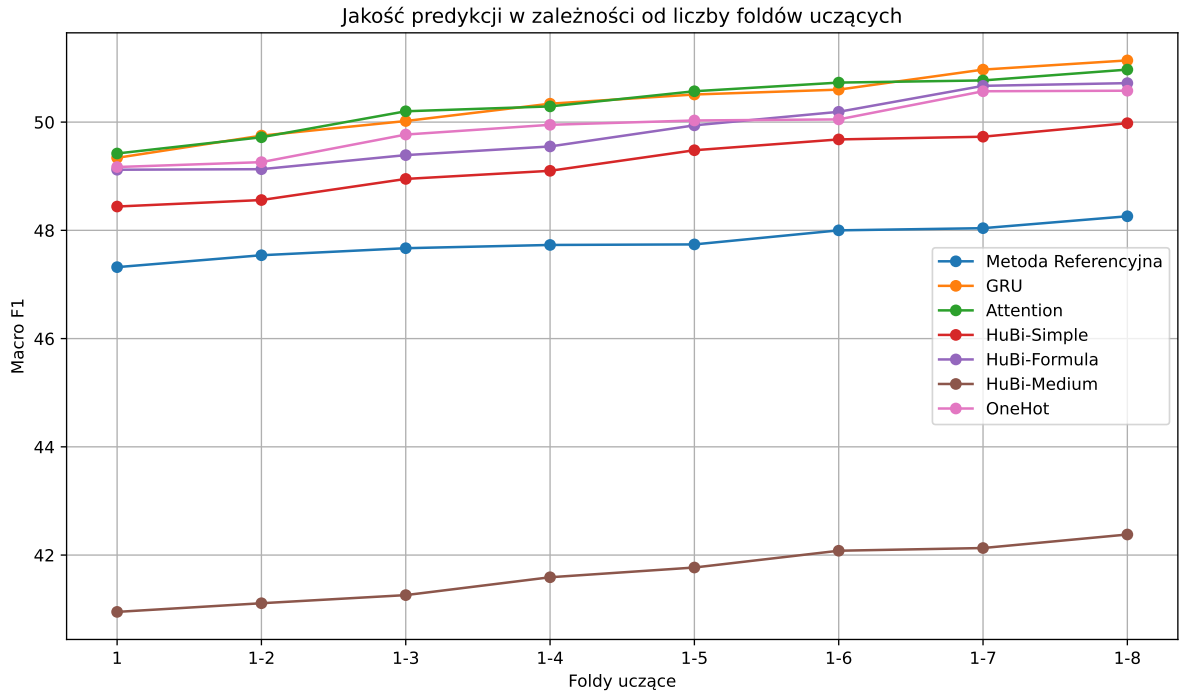
Rysunek 43: Wartości miary F1 macro w zależności od liczby foldów uczących dla zbioru Cockamamie. (Źródło: Opracowanie własne).



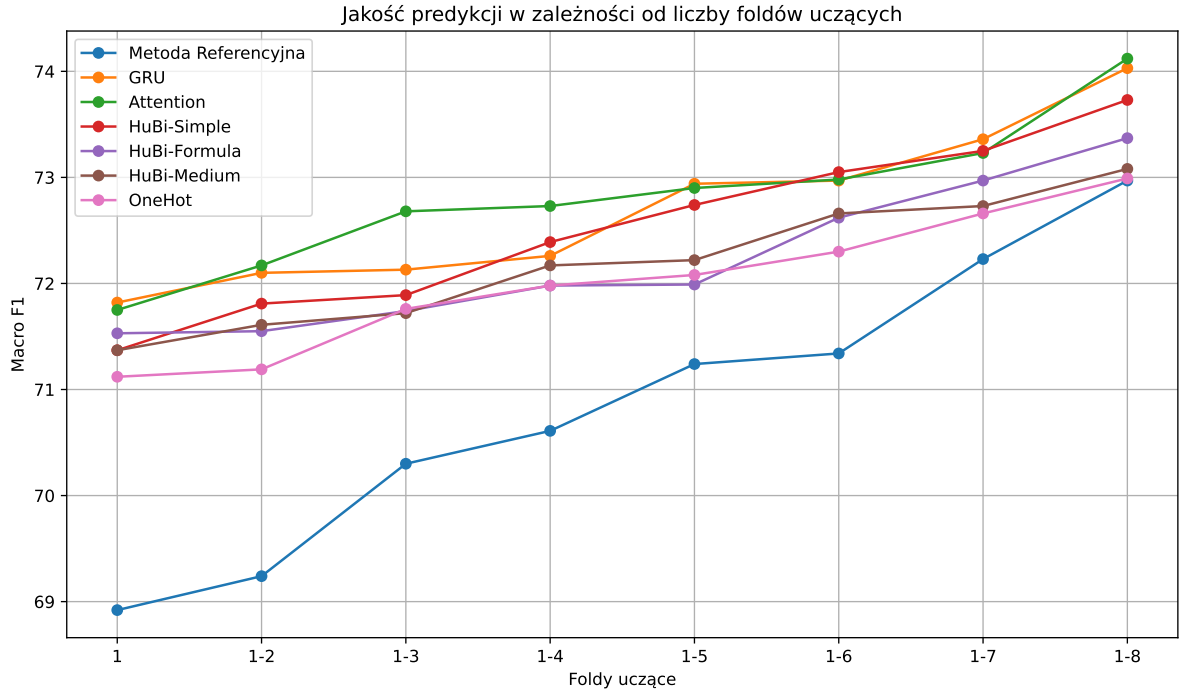
Rysunek 44: Wartości miary precyzji w zależności od liczby foldów uczących dla zbioru Cockamamie. (Źródło: Opracowanie własne).



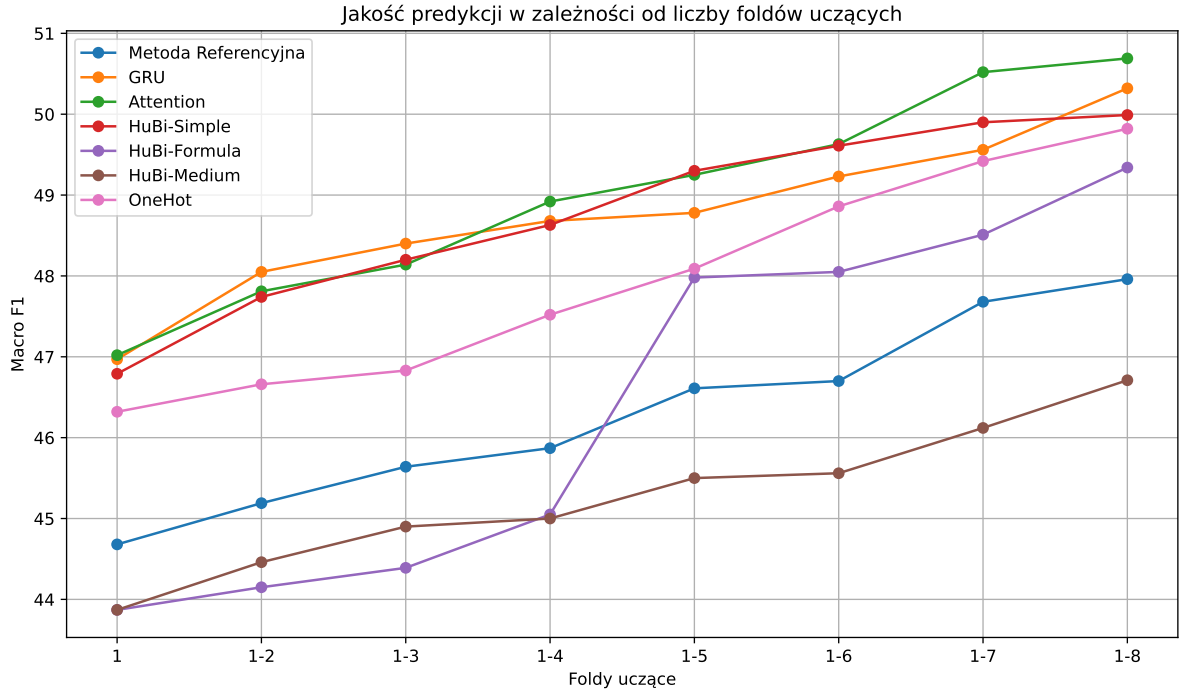
Rysunek 45: Wartości miary kompletności w zależności od liczby foldów uczących dla zbioru Cockamamie. (Źródło: Opracowanie własne).



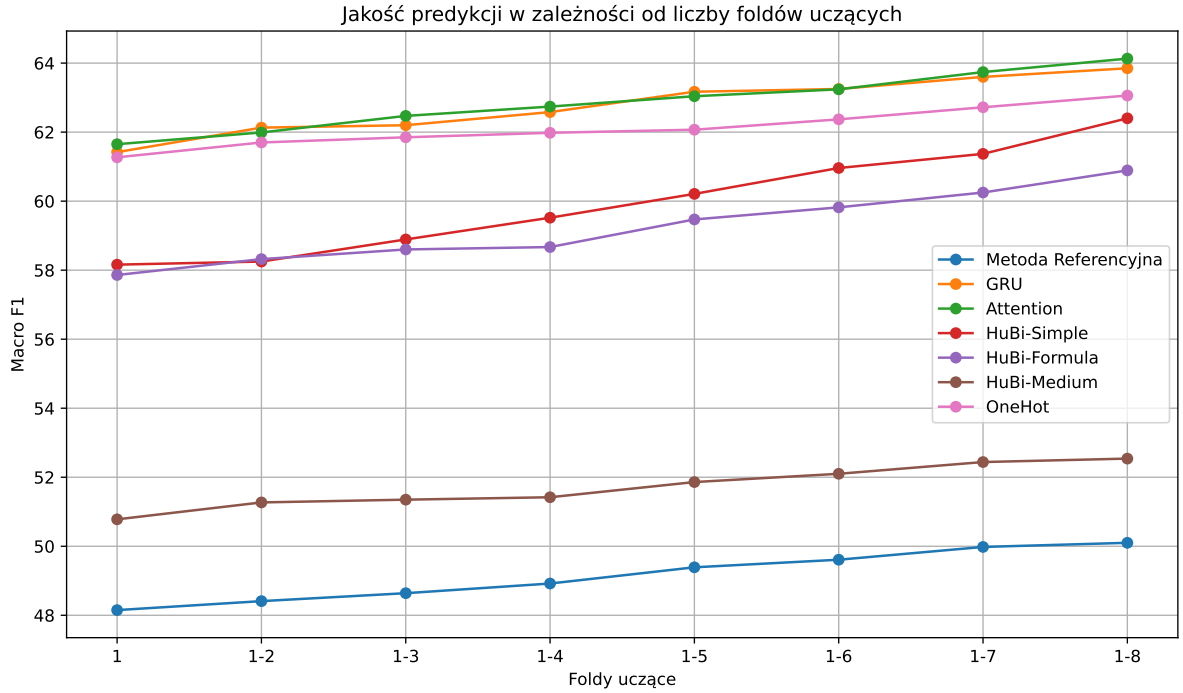
Rysunek 46: Wartości miary Macro F1 w zależności od liczby foldów uczących dla zbioru Humicroedit. (Źródło: Opracowanie własne).



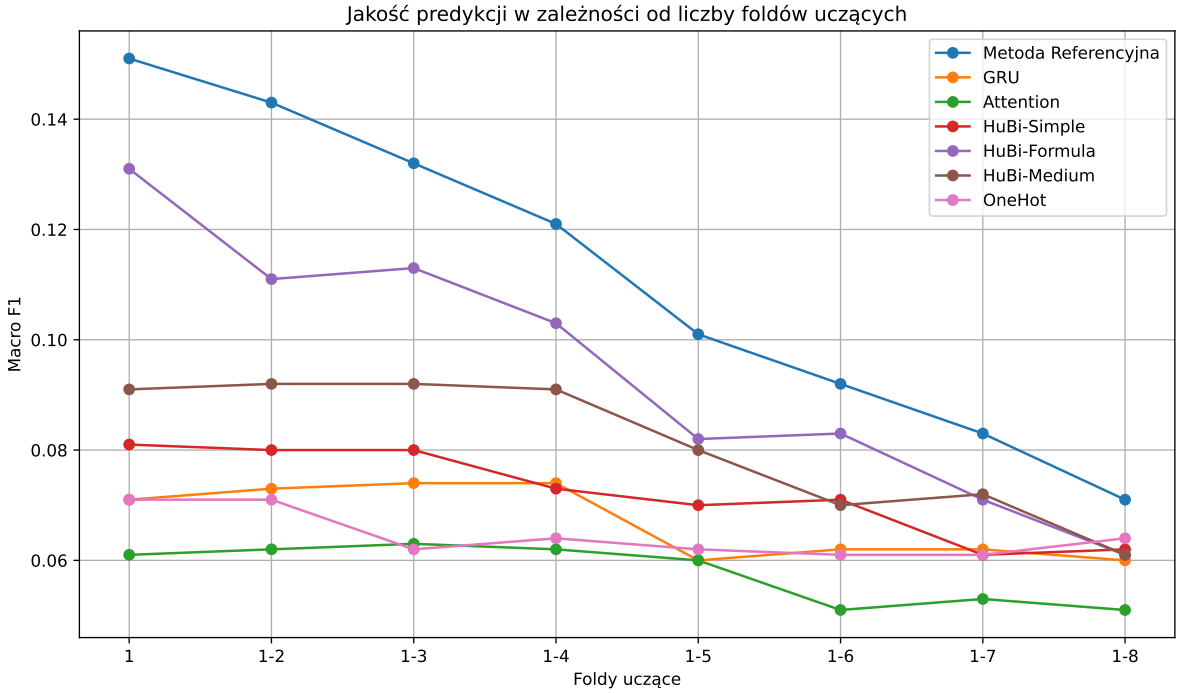
Rysunek 47: Wartości miary Macro F1 w zależności od liczby foldów uczących dla zbioru Humor. (Źródło: Opracowanie własne).



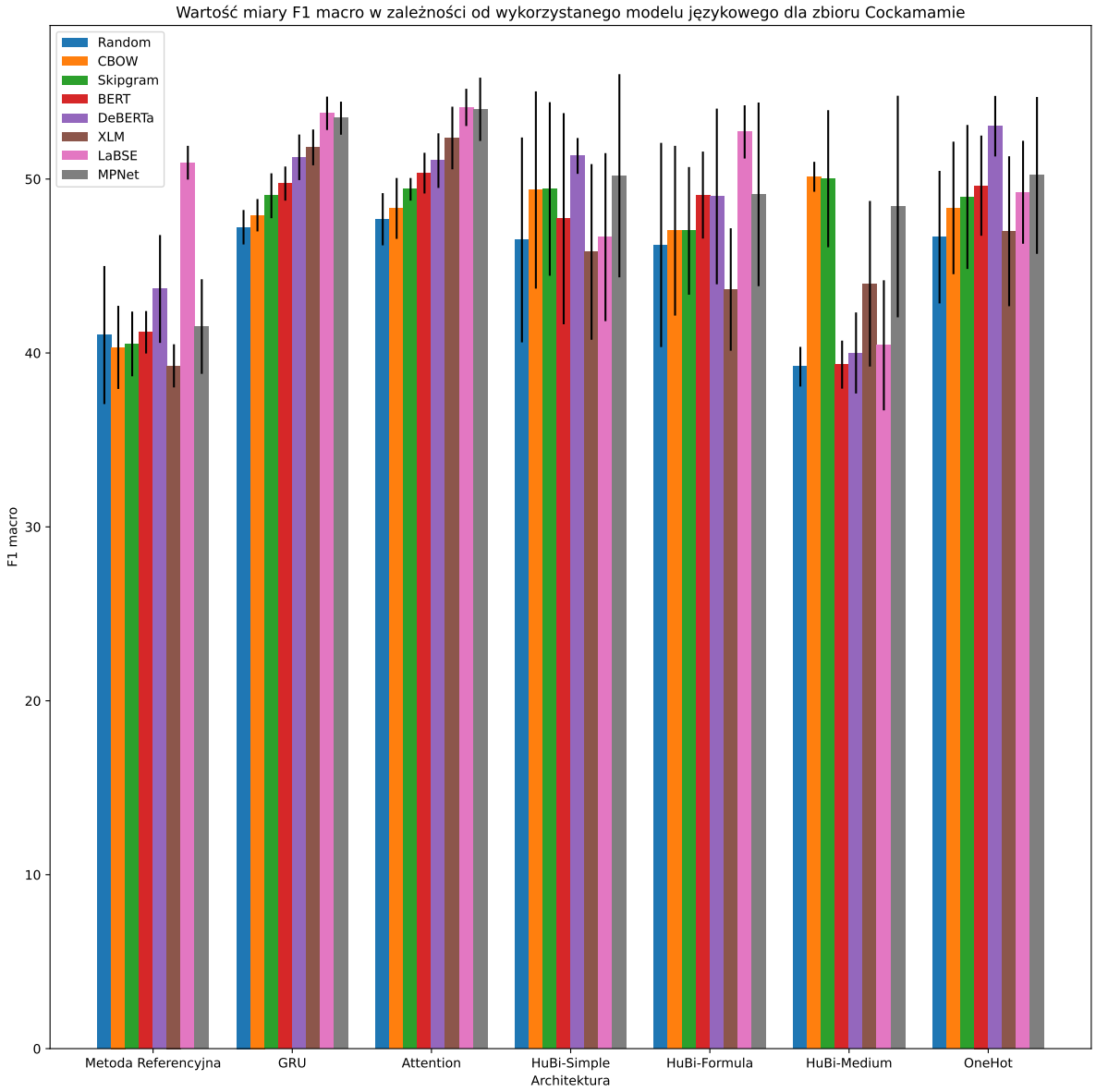
Rysunek 48: Wartości miary Macro F1 w zależności od liczby foldów uczących dla zbioru Doccano 1. (Źródło: Opracowanie własne).



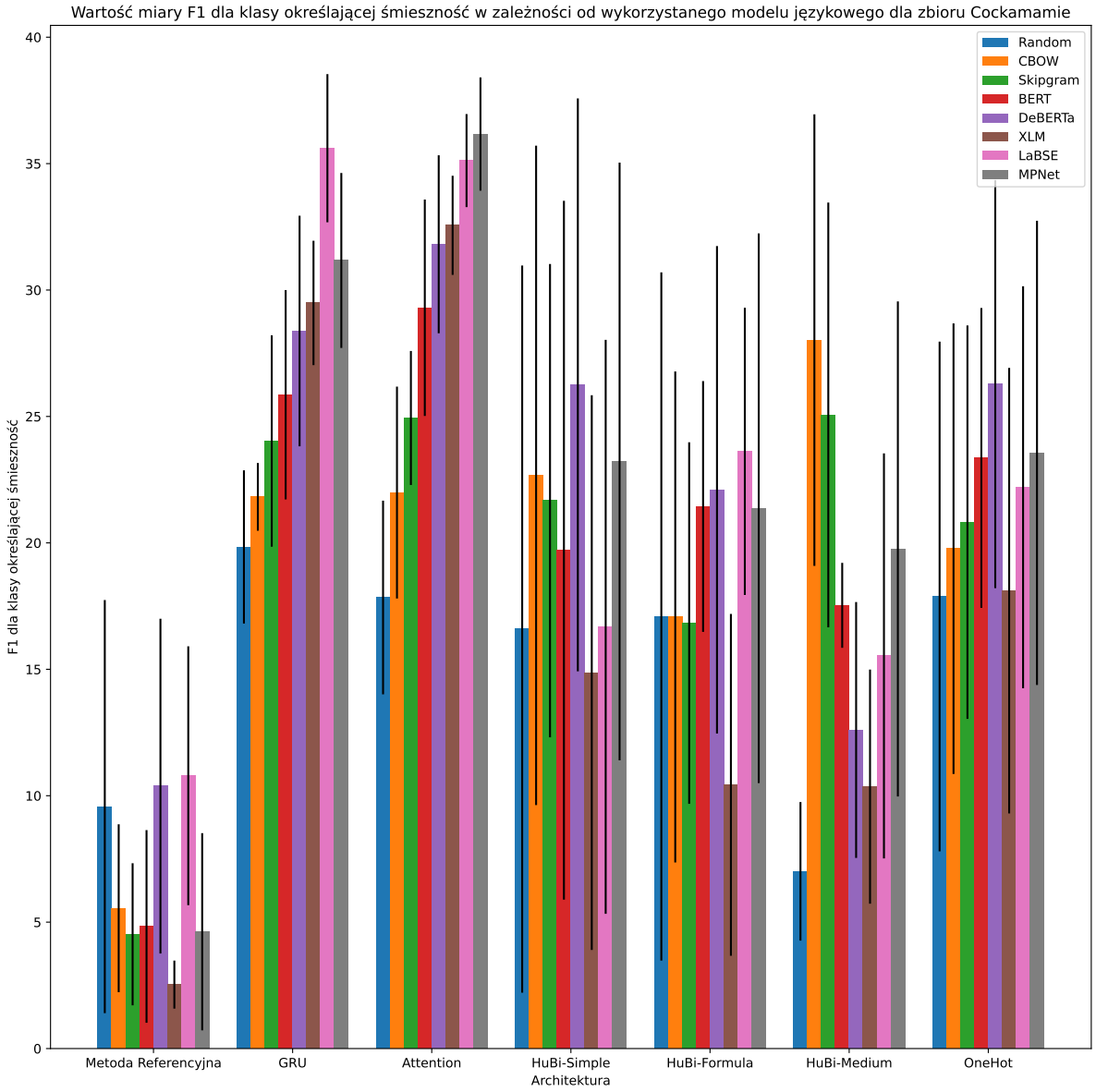
Rysunek 49: Wartości miary Macro F1 w zależności od liczby foldów uczących dla zbioru Doccano 2. (Źródło: Opracowanie własne).



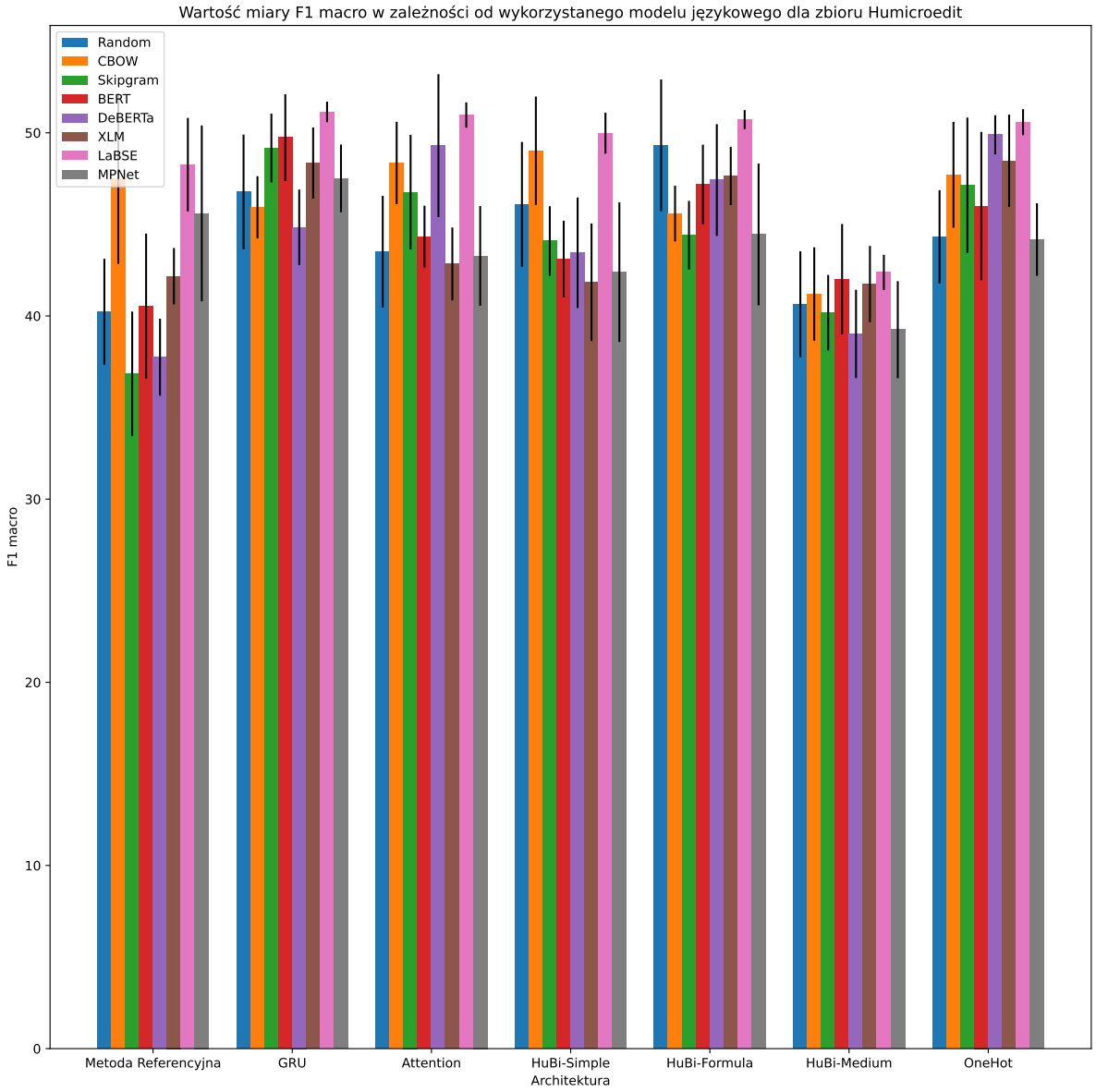
Rysunek 50: Wartości miary Mean Squared Error w zależności od liczby foldów uczących dla zbioru Jester. (Źródło: Opracowanie własne).



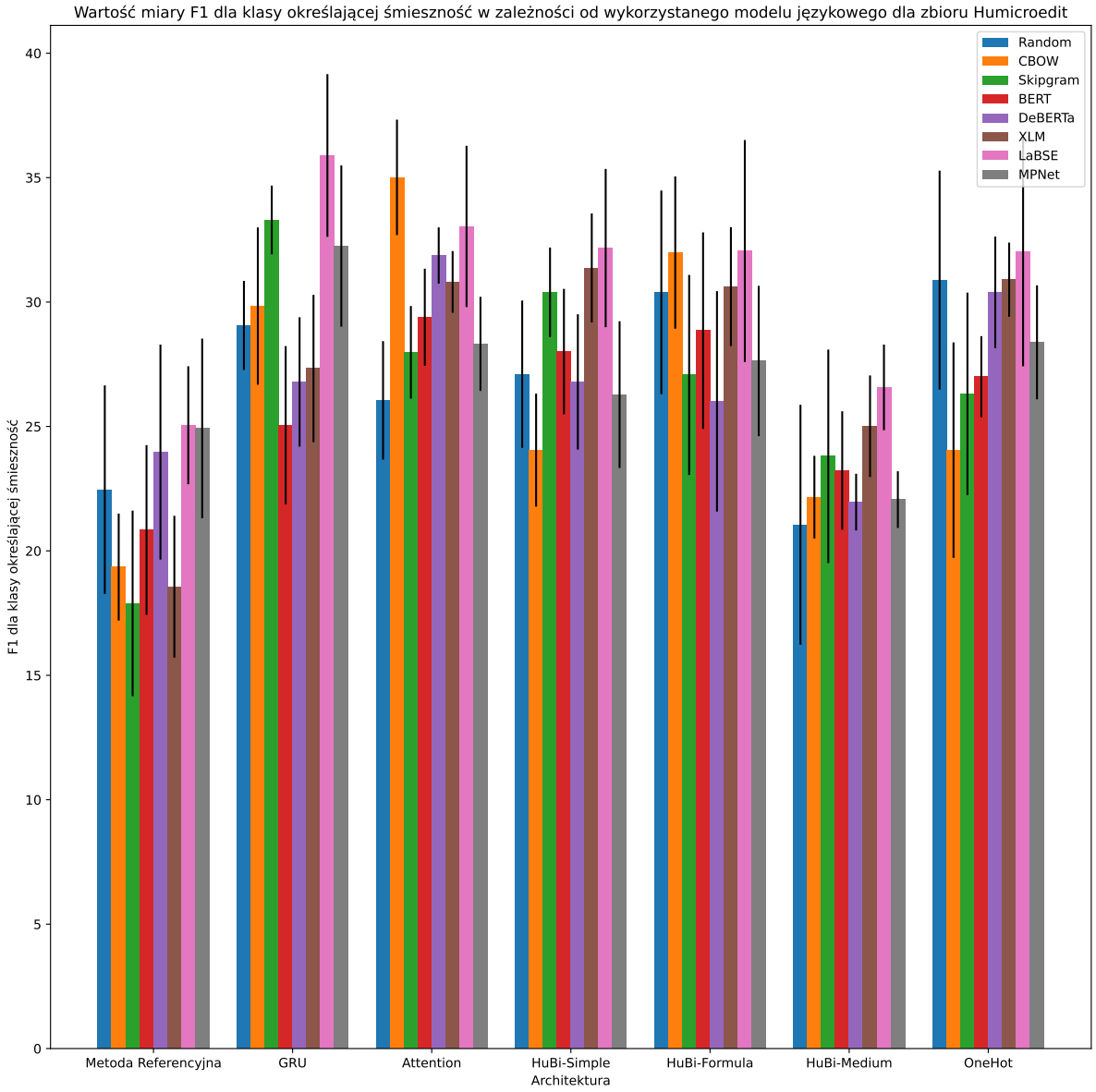
Rysunek 51: Wartości miary F1 macro w zależności od wykorzystanego modelu językowego dla zbioru Cockamamie. (Źródło: Opracowanie własne).



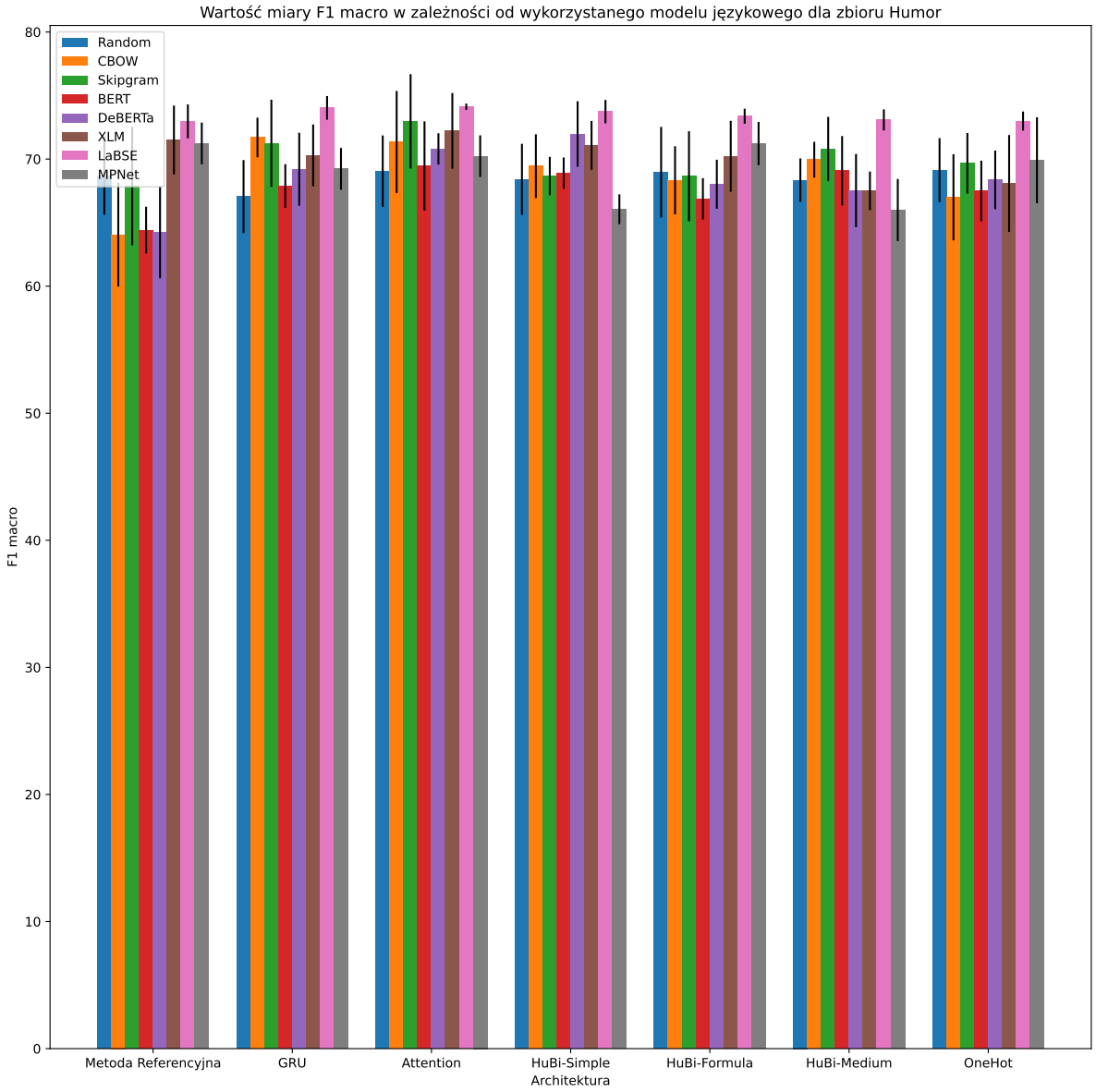
Rysunek 52: Wartości miary F1 dla klasy oznaczającej śmieszność w zależności od wykorzystanego modelu językowego dla zbioru Cockamamie. (Źródło: Opracowanie własne).



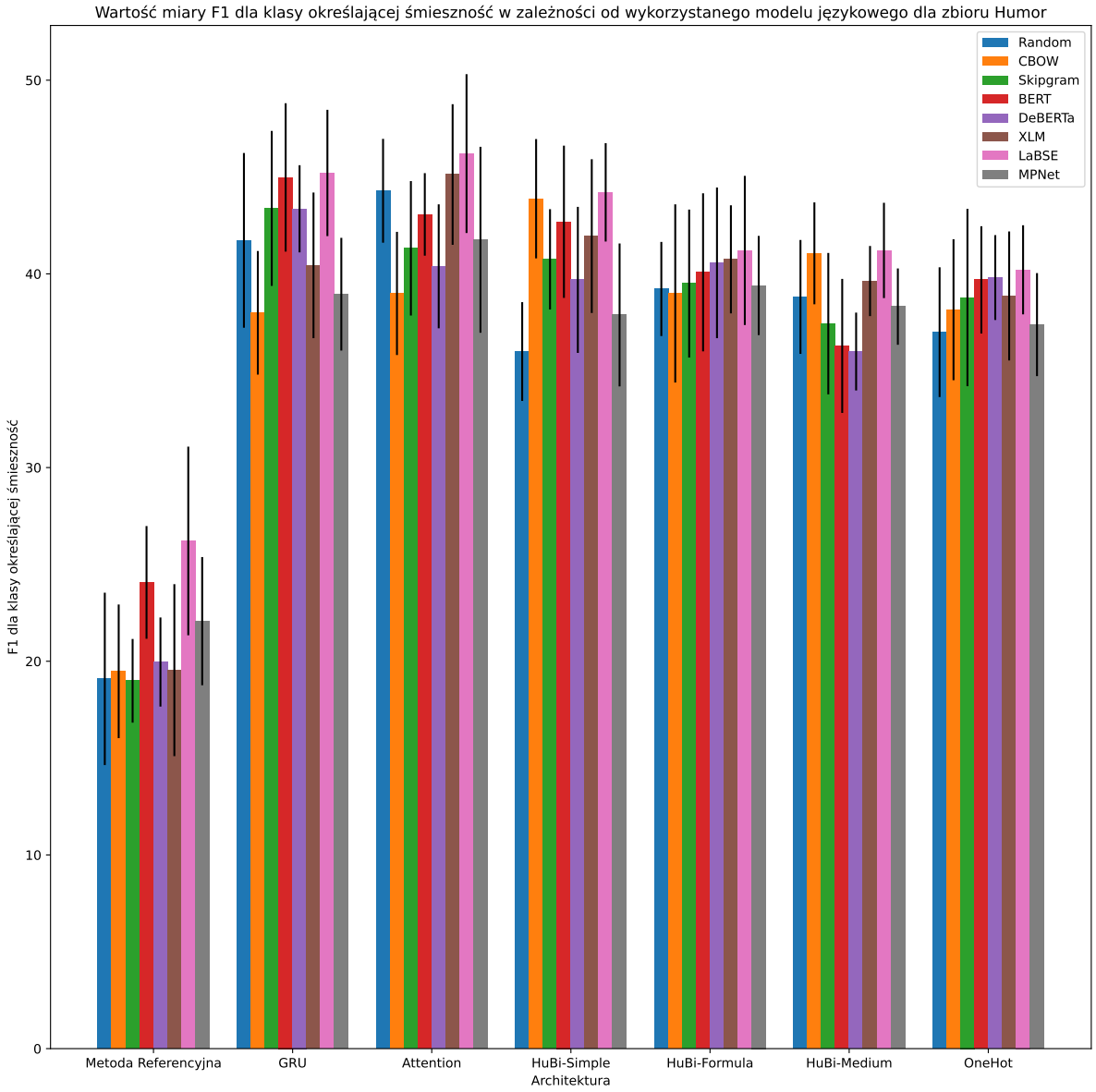
Rysunek 53: Wartości miary F1 macro w zależności od wykorzystanego modelu językowego dla zbioru Humicroedit. (Źródło: Opracowanie własne).



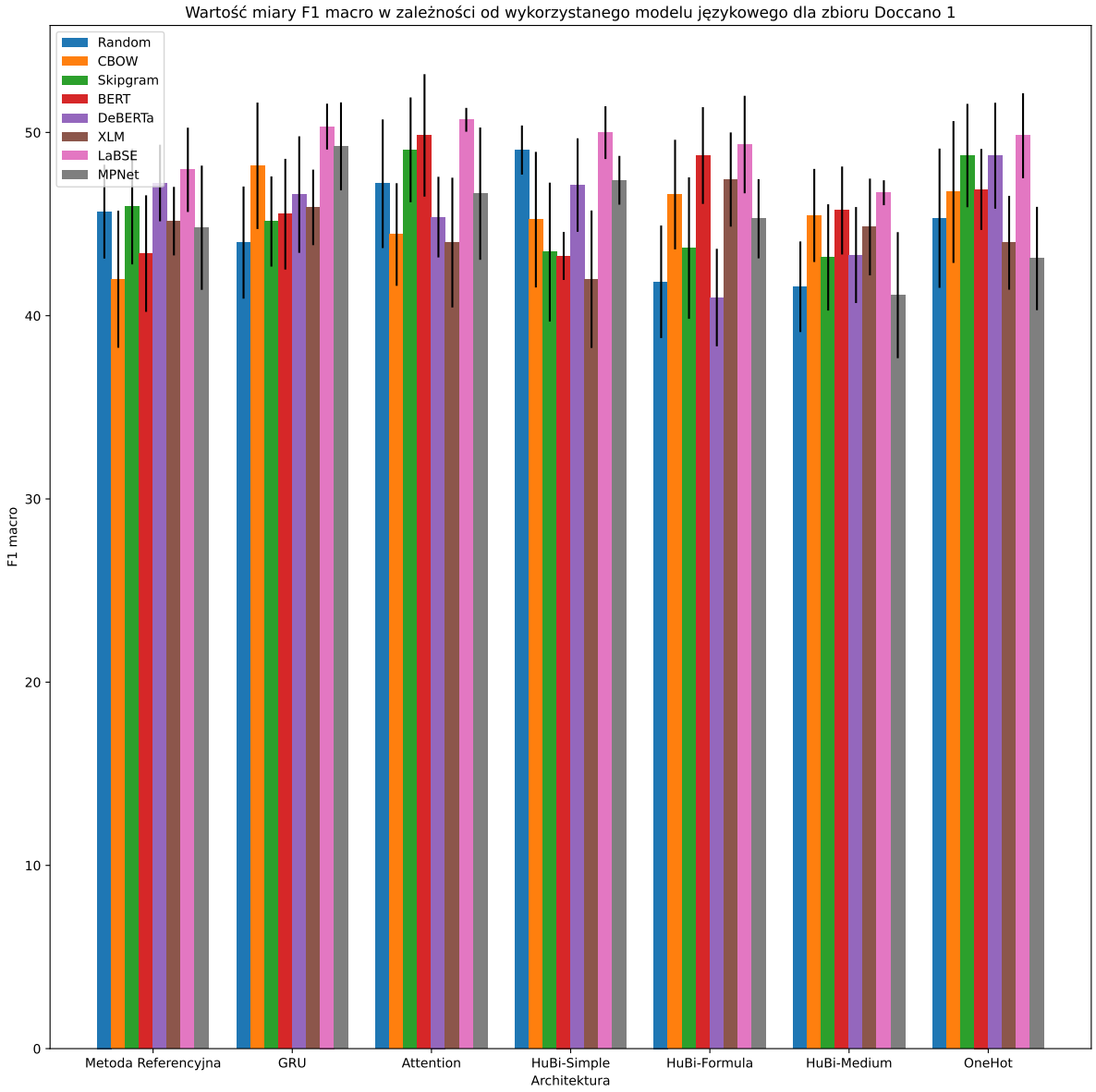
Rysunek 54: Wartości miary F1 dla klasy oznaczającej śmieszność w zależności od wykorzystanego modelu językowego dla zbioru Humicroedit. (Źródło: Opracowanie własne).



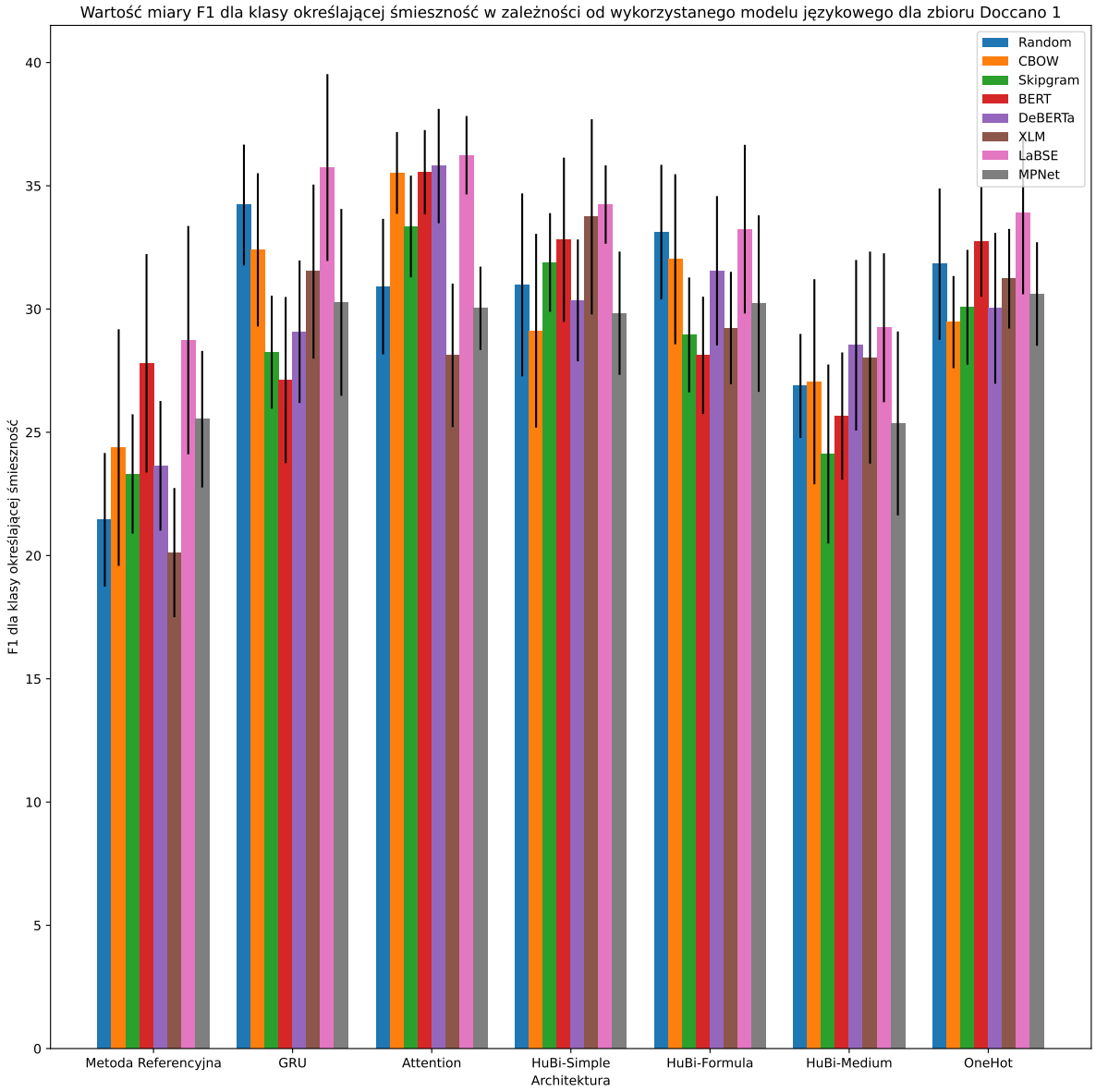
Rysunek 55: Wartości miary F1 macro w zależności od wykorzystanego modelu językowego dla zbioru Humor. (Źródło: Opracowanie własne).



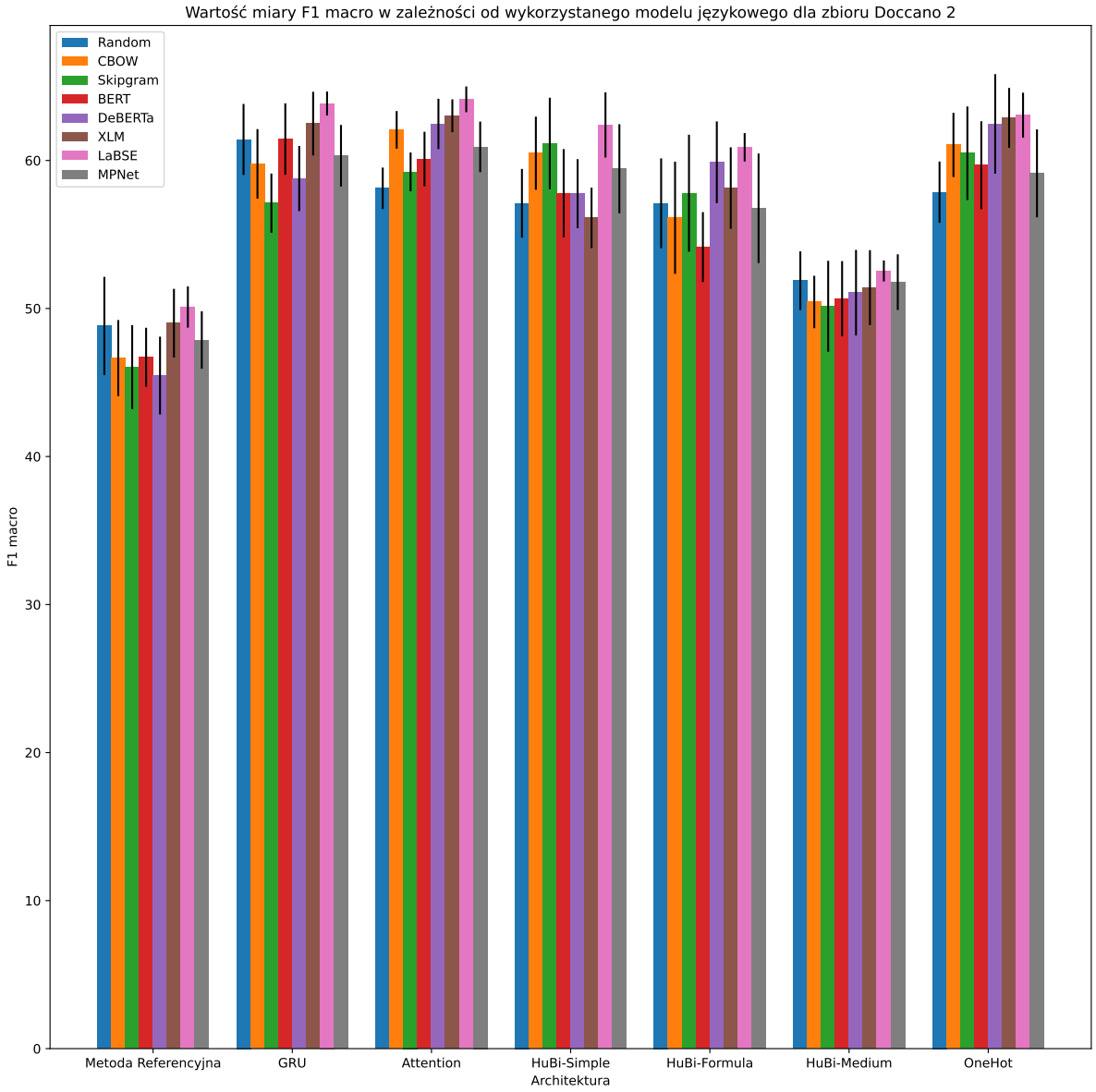
Rysunek 56: Wartości miary F1 dla klasy oznaczającej śmieszność w zależności od wykorzystanego modelu językowego dla zbioru Humor. (Źródło: Opracowanie własne).



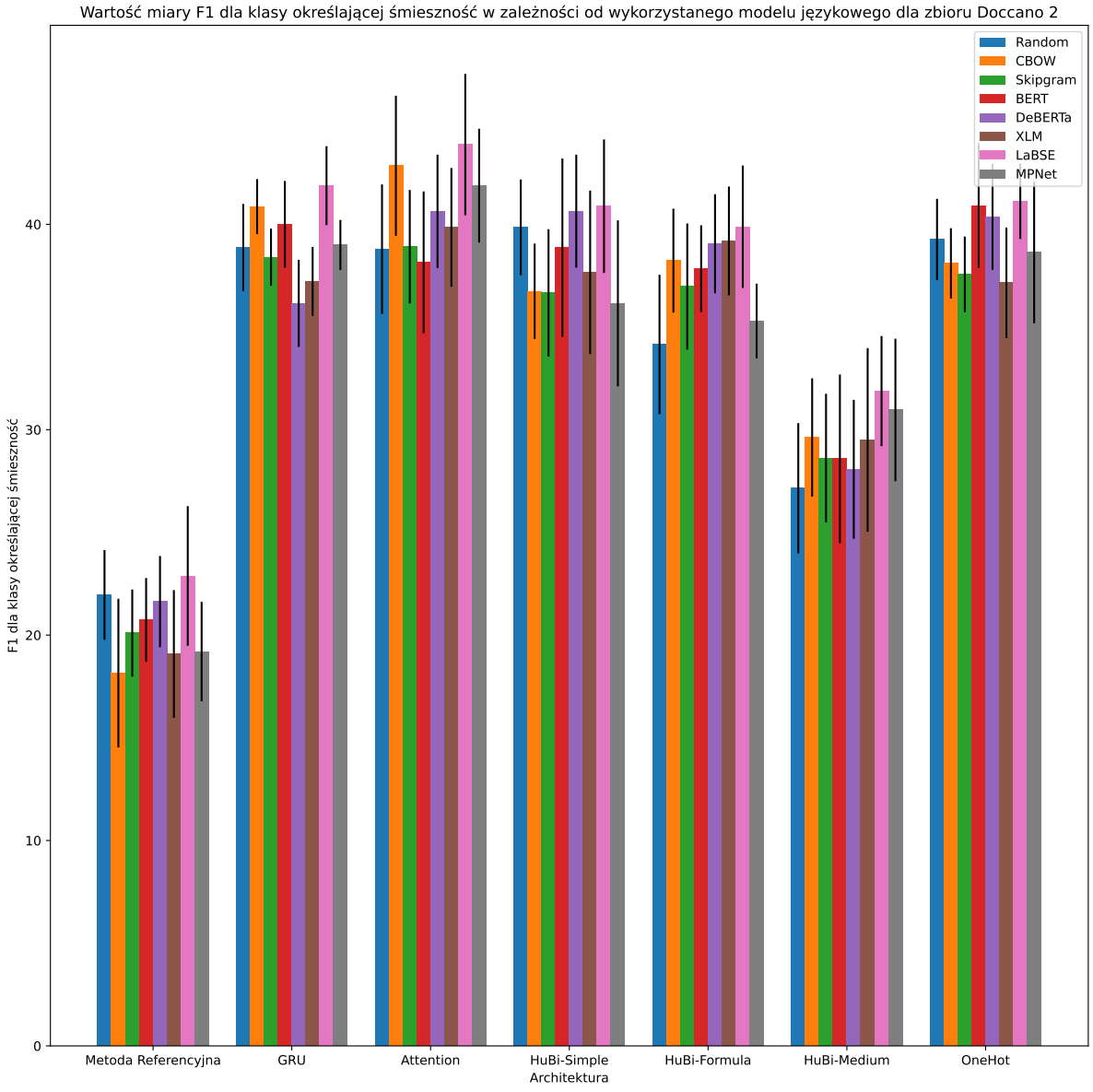
Rysunek 57: Wartości miary F1 macro w zależności od wykorzystanego modelu językowego dla zbioru Doccano 1. (Źródło: Opracowanie własne).



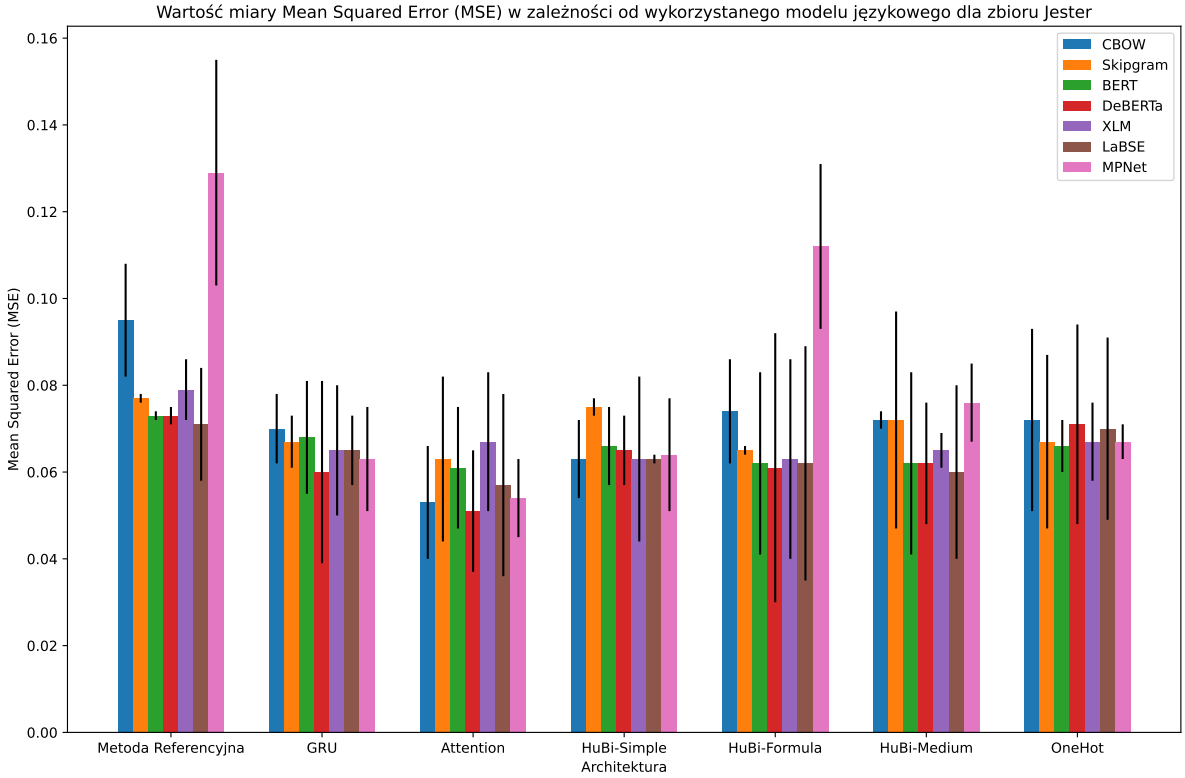
Rysunek 58: Wartości miary F1 dla klasy oznaczającej śmieszność w zależności od wykorzystanego modelu językowego dla zbioru Doccano 1. (Źródło: Opracowanie własne).



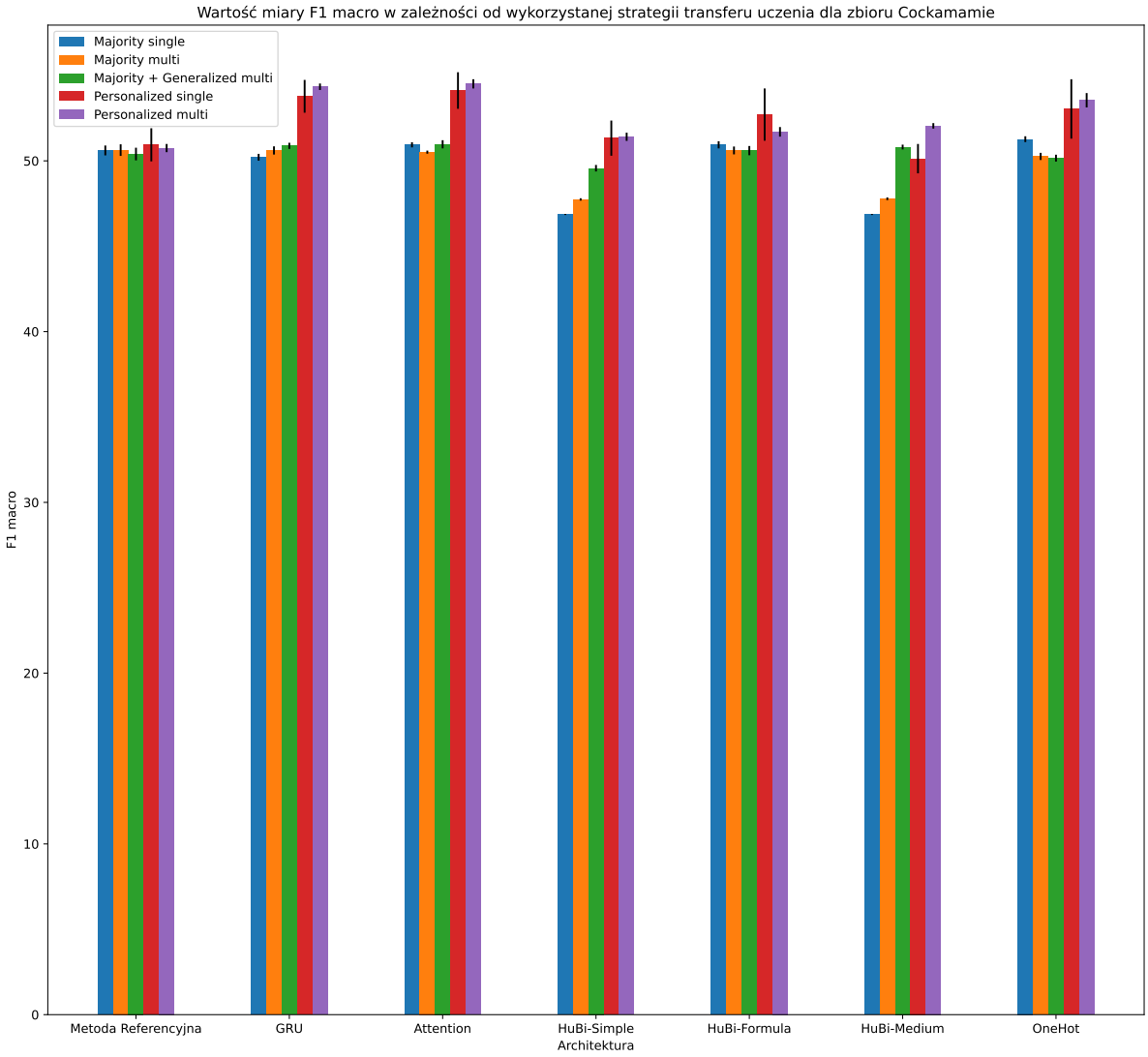
Rysunek 59: Wartości miary F1 macro w zależności od wykorzystanego modelu językowego dla zbioru Doccano 2. (Źródło: Opracowanie własne).



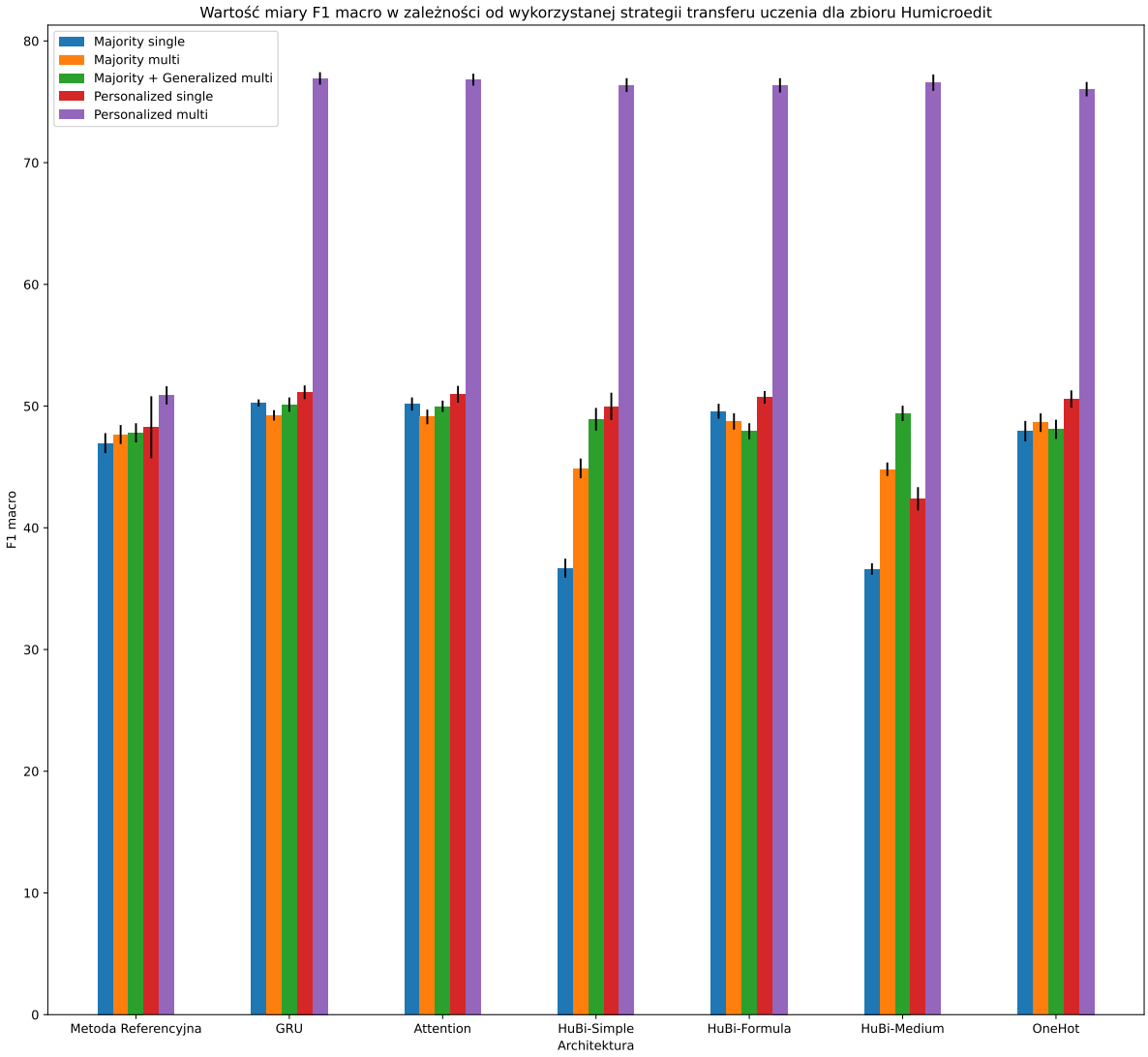
Rysunek 60: Wartości miary F1 dla klasy oznaczającej śmieszność w zależności od wykorzystanego modelu językowego dla zbioru Doccano 2. (Źródło: Opracowanie własne).



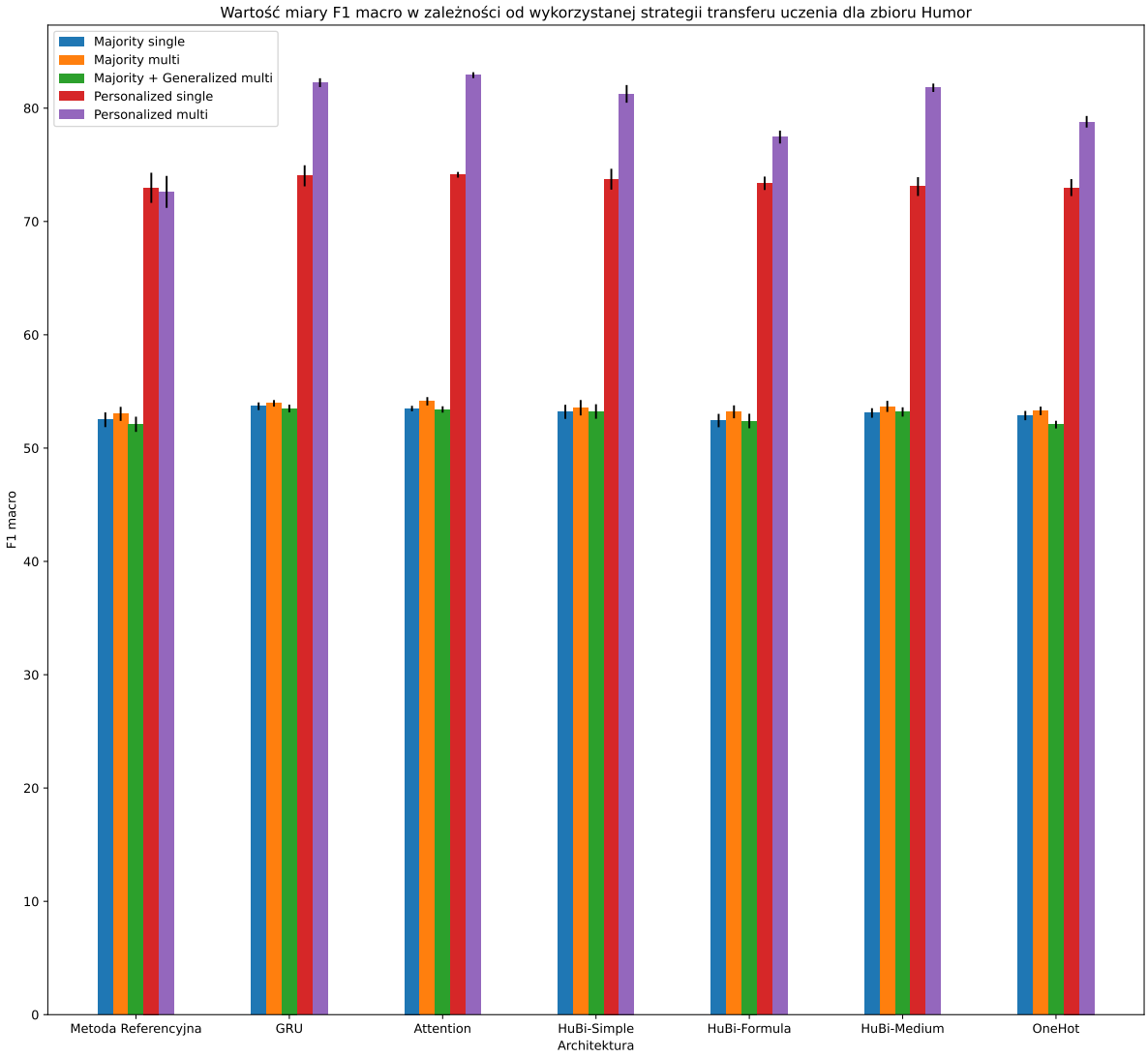
Rysunek 61: Wartości miary Mean Squared Error (MSE) w zależności od wykorzystanego modelu językowego dla zbioru Jester. (Źródło: Opracowanie własne).



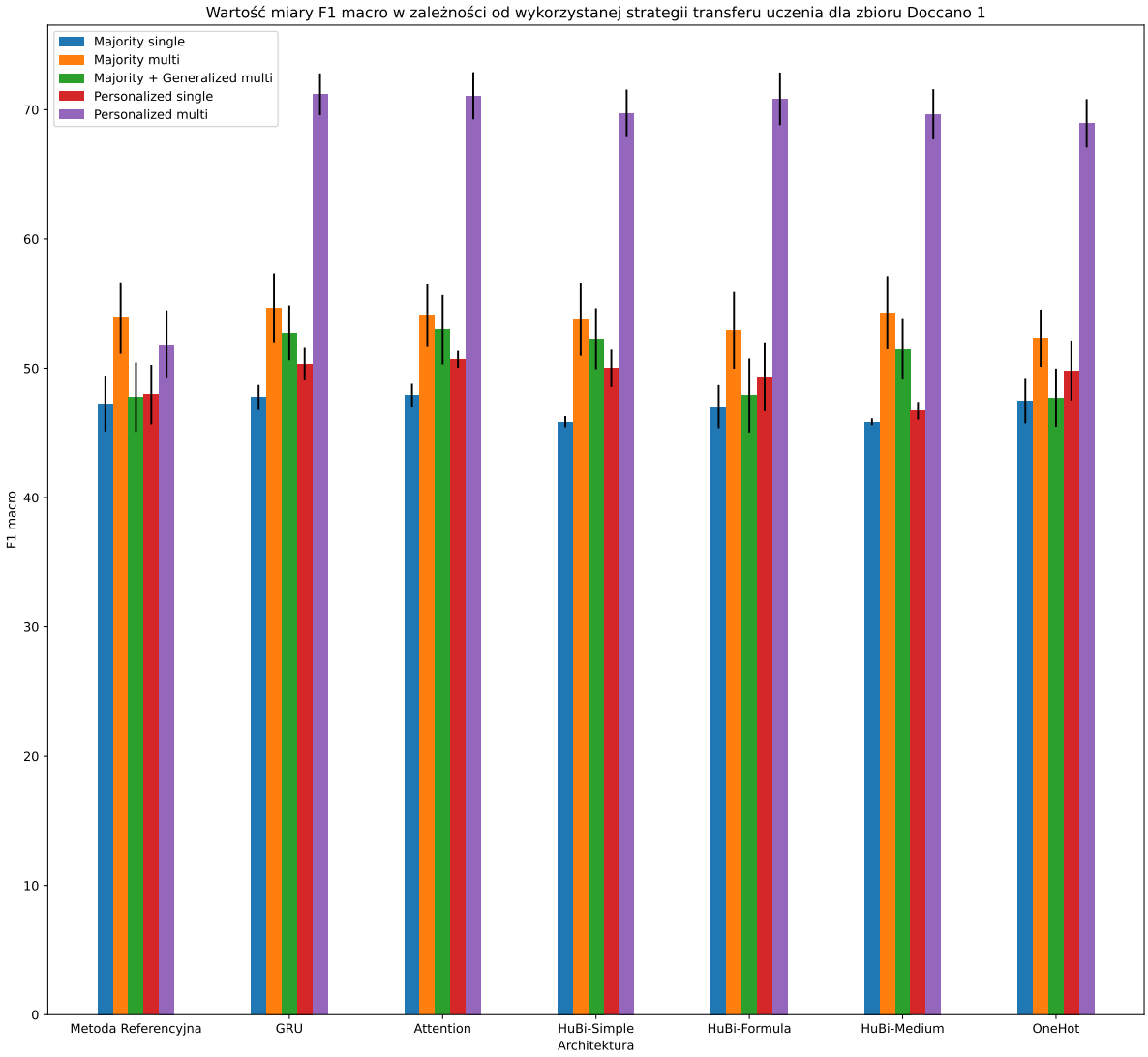
Rysunek 62: Wartości miary F1 macro w zależności od zastosowanej strategii transferu uczenia dla zbioru Cockamamie.



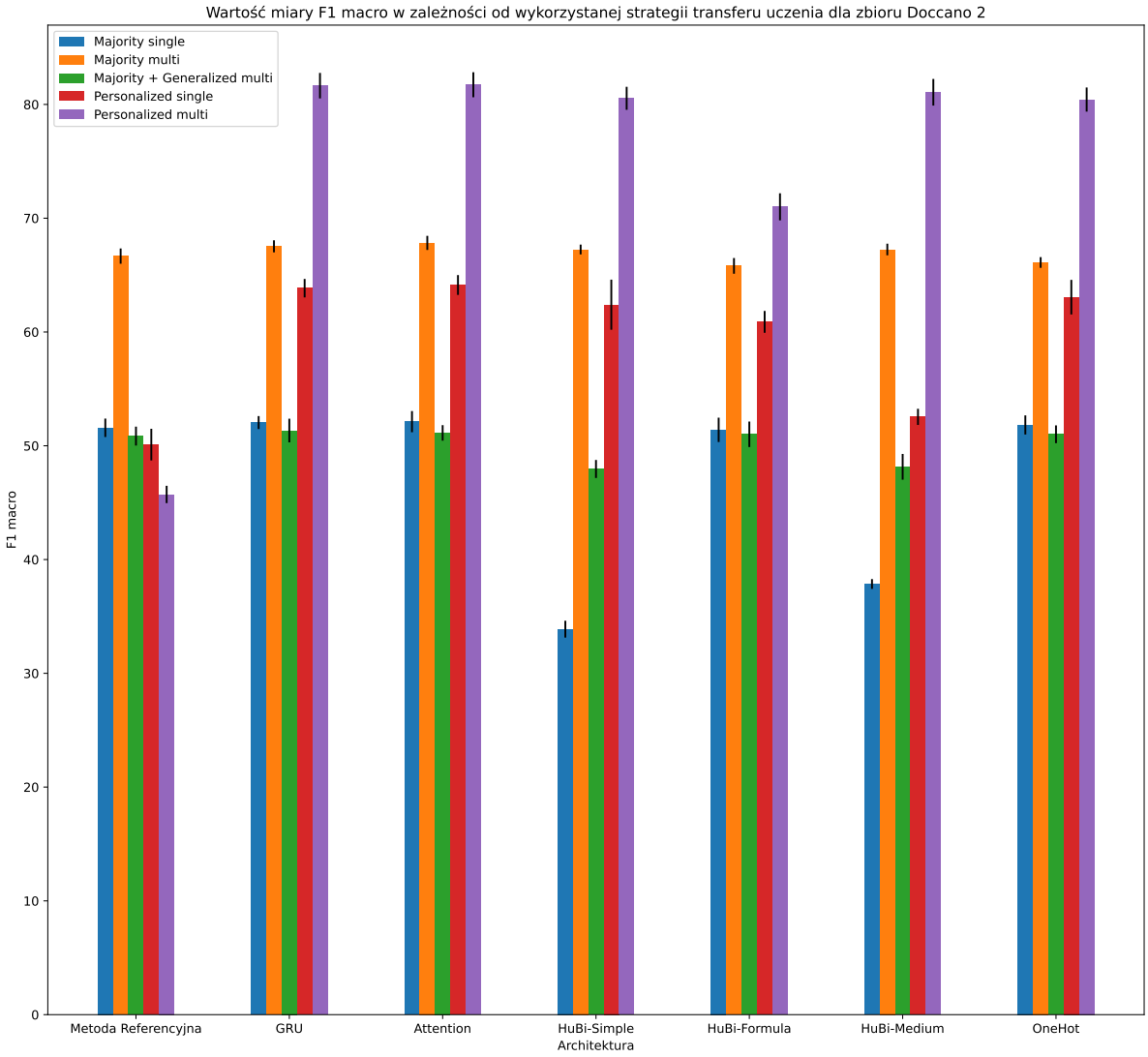
Rysunek 63: Wartości miary F1 macro w zależności od zastosowanej strategii transferu uczenia dla zbioru Humicroedit. (Źródło: Opracowanie własne).



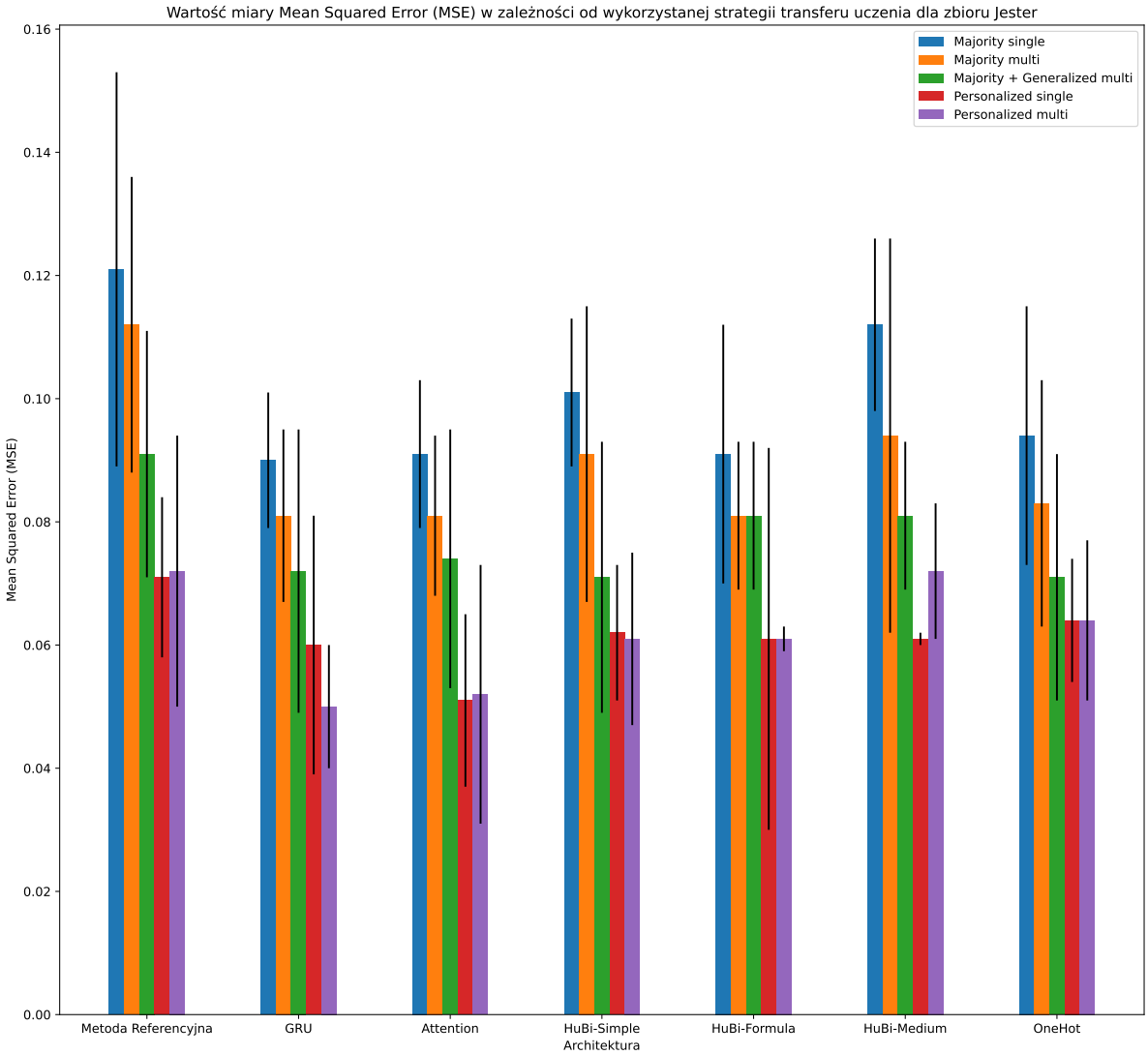
Rysunek 64: Wartości miary F1 macro w zależności od zastosowanej strategii transferu uczenia dla zbioru Humor. (Źródło: Opracowanie własne).



Rysunek 65: Wartości miary F1 macro w zależności od zastosowanej strategii transferu uczenia dla zbioru Doccano 1. (Źródło: Opracowanie własne).



Rysunek 66: Wartości miary F1 macro w zależności od zastosowanej strategii transferu uczenia dla zbioru Doccano 2. (Źródło: Opracowanie własne).



Rysunek 67: Wartości miary Mean Squared Error (MSE) w zależności od zastosowanej strategii transferu uczenia dla zbioru Jester. (Źródło: Opracowanie własne).

BIBLIOGRAFIA

Amir, Silvio, Byron C Wallace, Hao Lyu, Paula Carvalho i Mario J Silva (2016). "Modelling Context with User Embeddings for Sarcasm Detection in Social Media". W: *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, s. 167–177.

Annamoradnejad, Issa i Gohar Zoghi (2020). "Colbert: Using bert sentence embedding for humor detection". W: *arXiv preprint arXiv:2004.12765*.

Attardo, Salvatore (1994). "Linguistic theories of humor (Vol. 1)". W: *Mouton de Gruyter, Berlin & New York*.

Attardo, Salvatore i Victor Raskin (1991). "Script theory revis (it) ed: Joke similarity and joke representation model". W.

Bartlett, Maurice Stevenson (1937). "Properties of sufficiency and statistical tests". W: *Proceedings of the Royal Society of London. Series A-Mathematical and Physical Sciences* 160.901, s. 268–282.

Bielaniewicz, Julita, Kamil Kanclerz, Piotr Miłkowski, Marcin Gruza, Konrad Karanowski, Przemysław Kazienko i Jan Kocoń (2022a). "Deep-sheep: Sense of humor extraction from embeddings in the personalized context". W: *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, s. 967–974.

Bielaniewicz, Julita i Przemysław Kazienko (2023). "From Generalized Laughter to Personalized Chuckles: Unleashing the Power of Data Fusion in Subjective Humor Detection". W: *2023 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, s. 716–725.

Bielaniewicz, Julita i Adrianna Kozierkiewicz (2020). "A Hybrid Method of Facial Expressions Recognition in the Pictures". W: *Computational Collective Intelligence: 12th International Conference, ICCCI 2020, Da Nang, Vietnam, November 30–December 3, 2020, Proceedings 12*. Springer, s. 581–590.

Bielaniewicz, Julita i in. (2022b). "Deep-SHEEP: Sense of Humor Extraction from Embeddings in the Personalized Context". W: *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, s. 967–974. DOI: 10.1109/ICDMW58026.2022.00125.

Bojanowski, Piotr, Edouard Grave, Armand Joulin i Tomas Mikolov (2017). "Enriching word vectors with subword information". W: *Transactions of the Association for Computational Linguistics* 5, s. 135–146.

Castro, Santiago i in. (2018). "A Crowd-Annotated Spanish Corpus for Humor Analysis". W: *Proc. of SocialNLP 2018*. ACL, s. 7–11.

Chakrabarty, Navoneel i in. (2019). "RBM based joke recommendation system and joke reader segmentation". W: *PRML*. Springer, s. 229–239.

Chen, Y. i J. Martin (2007). "The Effects of Humor on Cognitive Performance and Learning". W: *Journal of Educational Psychology* 99.3, s. 481–493. DOI: 10.1037/0022-0663.99.3.481.

Chiruzzo, Luis i in. (2020). "HAHA 2019 dataset: A corpus for humor analysis in Spanish". W: *Proceedings of the Twelfth Language Resources and Evaluation Conference*, s. 5106–5112.

— (2021). "Overview of haha at iberlef 2021: Detecting, rating and analyzing humor in spanish". W: *Procesamiento del Lenguaje Natural*, s. 257–268.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer i Veselin Stoyanov (2020). "Unsupervised Cross-lingual Representation Learning at Scale". W: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, s. 8440–8451.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee i Kristina Toutanova (czer. 2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". W: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, s. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://www.aclweb.org/anthology/N19-1423.

Duncan, W Jack (1985). "The superiority theory of humor at work: Joking relationships as indicators of formal and informal status patterns in small, task-oriented groups". W: *Small Group Behavior* 16.4, s. 556–564.

Dunn, Olive Jean (1961). "Multiple comparisons among means". W: *Journal of the American statistical association* 56.293, s. 52–64.

Engelthaler, Tomas i in. (2018). "Humor norms for 4,997 English words". W: *Behavior research methods* 50.3, s. 1116–1124.

Feng, Fangxiaoyu i in. (2020). "Language-agnostic bert sentence embedding". W: *arXiv preprint arXiv:2007.01852*.

Freud, Sigmund (1971). *Der Witz und seine Beziehung zum Unbewussten*. Fischer.

Gervais, David i David K. Wilson (2005). "The Evolution and Functions of Laughter and Humor: A Synthetic Approach". W: *Quarterly Review of Biology* 80.4, s. 395–430. DOI: 10.1086/508789.

Goldberg, Kenneth i in. (2001). "Eigentaste: A Constant Time Collaborative Filtering Algorithm". W: *Information Retrieval* 4, s. 133–151. DOI: 10.1023/A:1011419012209.

Graham, Elizabeth E, Michael J Papa i Gordon P Brooks (1992). "Functions of humor in conversation: Conceptualization and measurement". W: *Western Journal of Communication (Includes Communication Reports)* 56.2, s. 161–183.

Hay, Jennifer (1995). "Gender and humour: Beyond a joke". W: *Unpublished Master's thesis, Victoria University of Wellington, Wellington, New Zealand*.

He, Pengcheng, Xiaodong Liu, Jianfeng Gao i Weizhu Chen (2020). "DeBERTa: Decoding-enhanced BERT with Disentangled Attention". W: *arXiv preprint arXiv:2006.03654*.

Hochreiter, Sepp i Jürgen Schmidhuber (1997). "Long Short-Term Memory". W: *Neural Computation* 9.8, s. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735.

Hofmann, Jennifer i in. (2020). "Gender differences in humor-related traits, humor appreciation, production, comprehension,(neural) responses, use, and correlates: A systematic review". W: *Current Psychology*, s. 1–14.

Hossain, Nabil, John Krumm, Michael Gamon i Henry Kautz (grud. 2020). "SemEval-2020 Task 7: Assessing Humor in Edited News Headlines". W: *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. International Committee for Computational Linguistics, s. 746–758. DOI: 10.18653/v1/2020.semeval-1.98. URL: https://aclanthology.org/2020.semeval-1.98.

Hossain, Nabil i in. (2019). ""President Vows to Cut <Taxes> Hair": Dataset and Analysis of Creative Text Editing for Humorous Headlines". W: *Proc. of the NAACL: Human Language Technologies*. ACL, s. 133–142. DOI: 10.18653/v1/N19-1012.

Jiang, Wen i Huiming Hou (2019). "Exploring the Effectiveness of Humor in Enhancing Learning and Memory". W: *Educational Psychology Review* 31.1, s. 115–133. DOI: 10.1007/s10648-018-9448-9.

Kanclerz, Kamil, Julita Bielaniewicz, Marcin Gruza, Jan Kocoń, Stanisław Woźniak i Przemysław Kazienko (2023a). "Towards Model-Based Data Acquisition for Subjective Multi-Task NLP Problems". W: *2023 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, s. 726–735.

Kanclerz, Kamil, Konrad Karanowski, Julita Bielaniewicz, Marcin Gruza, Piotr Miłkowski, Jan Kocoń i Przemyslaw Kazienko (2023b). "PALS: Personalized Active Learning for Subjective Tasks in NLP". W: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, s. 13326–13341.

Kanclerz, Kamil i in. (2020). "Cross-lingual deep neural transfer learning in sentiment analysis". W: *Procedia Computer Science* 176, s. 128–137.

— (2021). "Controversy and conformity: from generalized to personalized aggressiveness detection". W: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, s. 5915–5926.

Kanclerz, Kamil i in. (2022). "What if ground truth is subjective? personalized deep neural hate speech detection". W: *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP@ LREC2022*, s. 37–45.

Kazienko, Przemysław, Julita Bielaniewicz, Marcin Gruza, Kamil Kanclerz, Konrad Karanowski, Piotr Miłkowski i Jan Kocoń (2023). "Human-centered neural reasoning for subjective content processing: Hate speech, emotions, and humor". W: *Information Fusion* 94, s. 43–65.

Kiddon, Chloe i Yuriy Brun (2011). "That's what she said: double entendre identification". W: *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, s. 89–94.

Kocoń, Jan, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz i in. (2023). "ChatGPT: Jack of all trades, master of none". W: *Information Fusion*, s. 101861.

Kocoń, Jan, Alicja Figas, Marcin Gruza, Daria Puchalska, Tomasz Kajdanowicz i Przemysław Kazienko (2021a). "Offensive, aggressive, and hate speech analysis: From data-centric to human-centered approach". W: *Information Processing & Management* 58.5, s. 102643.

Kocoń, Jan, Marcin Gruza, Julita Bielaniewicz, Damian Grimling, Kamil Kanclerz, Piotr Miłkowski i Przemysław Kazienko (2021b). "Learning personal human biases and representations for subjective tasks in natural language processing". W: *2021 IEEE International Conference on Data Mining (ICDM)*. IEEE, s. 1168–1173.

Koestler, Arthur (1964). "The act of creation". W: *London: Hutchinson*.

Kulka, Tomáš (2007). "The incongruity of incongruity theories of humor". W: *Organon F* 14.3, s. 320–333.

Lambert South, Andrea, Jessica Elton i Alison M Lietzenmayer (2022). "Communicating death with humor: Humor types and functions in death over dinner conversations". W: *Death studies* 46.4, s. 851–860.

Leist, Anja K i Daniela Müller (2013). "Humor types show different patterns of self-regulation, self-esteem, and well-being". W: *Journal of Happiness Studies* 14, s. 551–569.

Liu, Andrew, P. He, Q. Chen, O. Levy, Y. Zhang, W. Yih i J. Zhou (2019). "Fine-Tuned Language Models for Text Classification". W: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, s. 4776–4784. URL: https://www.aclweb.org/anthology/D19-1485/.

Mann, Henry B i Donald R Whitney (1947). "On a test of whether one of two random variables is stochastically larger than the other". W: *The annals of mathematical statistics*, s. 50–60.

Martin, Richard i James Ford (2018). "The Influence of Humor on Memory and Learning". W: *Journal of Cognitive Psychology* 30.2, s. 145–160. DOI: 10.1080/20445911.2017.1392184.

Meaney, J. A. i in. (sierp. 2021). "SemEval 2021 Task 7: HaHackathon, Detecting and Rating Humor and Offense". W: *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*. Online: Association for Computational Linguistics, s. 105–119.

Miao, Zhongchen i in. (2016). "Joint prediction of rating and popularity for cold-start item by sentinel user selection". W: *IEEE Access* 4.

Mieleszczenko-Kowszewicz, Wiktoria, Kamil Kanclerz, Julita Bielaniewicz, Marcin Oleksy, Marcin Gruza, Stanisław Woźniak, Ewa Dziecioł, Przemysław Kazienko i Jan Kocoń (2023). "Capturing Human Perspectives in NLP: Questionnaires, Annotations, and Biases". W: *The ECAI 2023 2nd Workshop on Perspectivist Approaches to NLP*. CEUR Workshop Proceedings.

Mihalcea, Rada i Stephen Pulman (2007). "Characterizing humour: An exploration of features in humorous texts". W: *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, s. 337–347.

Mihalcea, Rada i Carlo Strapparava (2005). "Making computers laugh: Investigations in automatic humor recognition". W: *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, s. 531–538.

Mireshghallah, Fatemehsadat, Vaishnavi Shrivastava, Milad Shokouhi, Taylor Berg-Kirkpatrick, Robert Sim i Dimitrios Dimitriadis (2021). "UserIdentifier: Implicit User Representations for Simple and Effective Personalized Sentiment Analysis". W: *arXiv:2110.00135*. arXiv: 2110.00135 [cs.LG].

Neuliep, James W (1991). "An examination of the content of high school teachers' humor in the classroom and the development of an inductively derived taxonomy of classroom humor". W: *Communication education* 40.4, s. 343–355.

Proyer, Klaus M. (2013). "The Role of Humor in Promoting Well-Being: A Review of Research". W: *European Journal of Personality* 27.4, s. 351–368. DOI: 10.1002/per.1884.

Purandare, Amruta i Diane Litman (2006). "Humor: Prosody analysis and automatic recognition for f* r* i* e* n* d* s". W: *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, s. 208–215.

Raskin, Victor (1979). "Semantic mechanisms of humor". W: *Annual Meeting of the Berkeley Linguistics Society*. T. 5, s. 325–335.

Raskin, Victor, Christian F Hempelmann i Julia M Taylor (2009). "How to understand and assess a theory: The evolution of the SSTH into the GTVH and now into the OSTH". W.

Shapiro, Samuel Sanford i Martin B Wilk (1965). "An analysis of variance test for normality (complete samples)". W: *Biometrika* 52.3-4, s. 591–611.

Shurcliff, Arthur (1968). "Judged humor, arousal, and the relief theory." W: *Journal of personality and social psychology* 8.4p1, s. 360.

Siddiqui, Ramsha (2019). *SARCASMANIA: Sarcasm Exposed*.

Song, Kaitao, Xu Tan, Tao Qin, Jianfeng Lu i Tie-Yan Liu (2020). "MPNet: Masked and Permuted Pre-training for Language Understanding". W: *Advances in Neural Information Processing Systems*. Red. H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan i H. Lin. T. 33. Curran Associates, Inc., s. 16857–16867. URL: https://proceedings.neurips.cc/paper/2020/file/c3a690be93aa602ee2dc0ccab5b7b67e-Paper.pdf.

Stock, Oliviero i Carlo Strapparava (2003). "Getting serious about the development of computational humor". W: *IJCAI*. T. 3, s. 59–64.

— (2006). "Automatic Production of Humorous Expressions for Catching the Attention and Remembering (in Trends and Controversies: Computational {H} umor)". W: *IEEE Intelligent Systems* 21.2, s. 64–67.

Student (1908). "The probable error of a mean". W: *Biometrika*, s. 1–25.

Tang, Duyu i in. (2015). "Learning semantic representations of users and products for document level sentiment classification". W: *ACL-IJCNLP*.

Taylor, Julia M i Lawrence J Mazlack (2004). "Computationally recognizing wordplay in jokes". W: *Proceedings of the Annual Meeting of the Cognitive Science Society*. T. 26. 26.

Veale, Tony (2004). "Incongruity in humor: Root cause or epiphenomenon?" W.

Vinton, Karen L (1989). "Humor in the workplace: It is more than telling jokes". W: *Small group behavior* 20.2, s. 151–166.

Westbury, Chris i in. (b.d.). "Telling the world's least funny jokes: On the quantification of humor as entropy". W: *Journal of Memory and Language* (). DOI: https://doi.org/10.1016/j.jml.2015.09.001.

Wilcoxon, Frank (1945). "Individual Comparisons by Ranking Methods". W: *Biometrics Bulletin* 1.6, s. 80–83. ISSN: 00994987. URL: http://www.jstor.org/stable/3001968 (term. wiz. 24. 09. 2024).

Woźniak, Stanisław, Bartłomiej Koptyra, Arkadiusz Janz, Przemysław Kazienko i Jan Kocoń (2024). "Personalized large language models". W: *arXiv preprint arXiv:2402.09269*.

Yang, Diyi i in. (2015). "Humor Recognition and Humor Anchor Extraction". W: *Proc. of the EMNLP*. ACL, s. 2367–2376. DOI: 10.18653/v1/D15-1284.

Yang, Qichuan i in. (2019). "End-to-End Personalized Humorous Response Generation in Untrimmed Multi-Role Dialogue System". W: *IEEE Access* 7, s. 94059–94071.

Zhang, Renxian i in. (2014). "Recognizing humor on twitter". W: *ACM CIKM*.